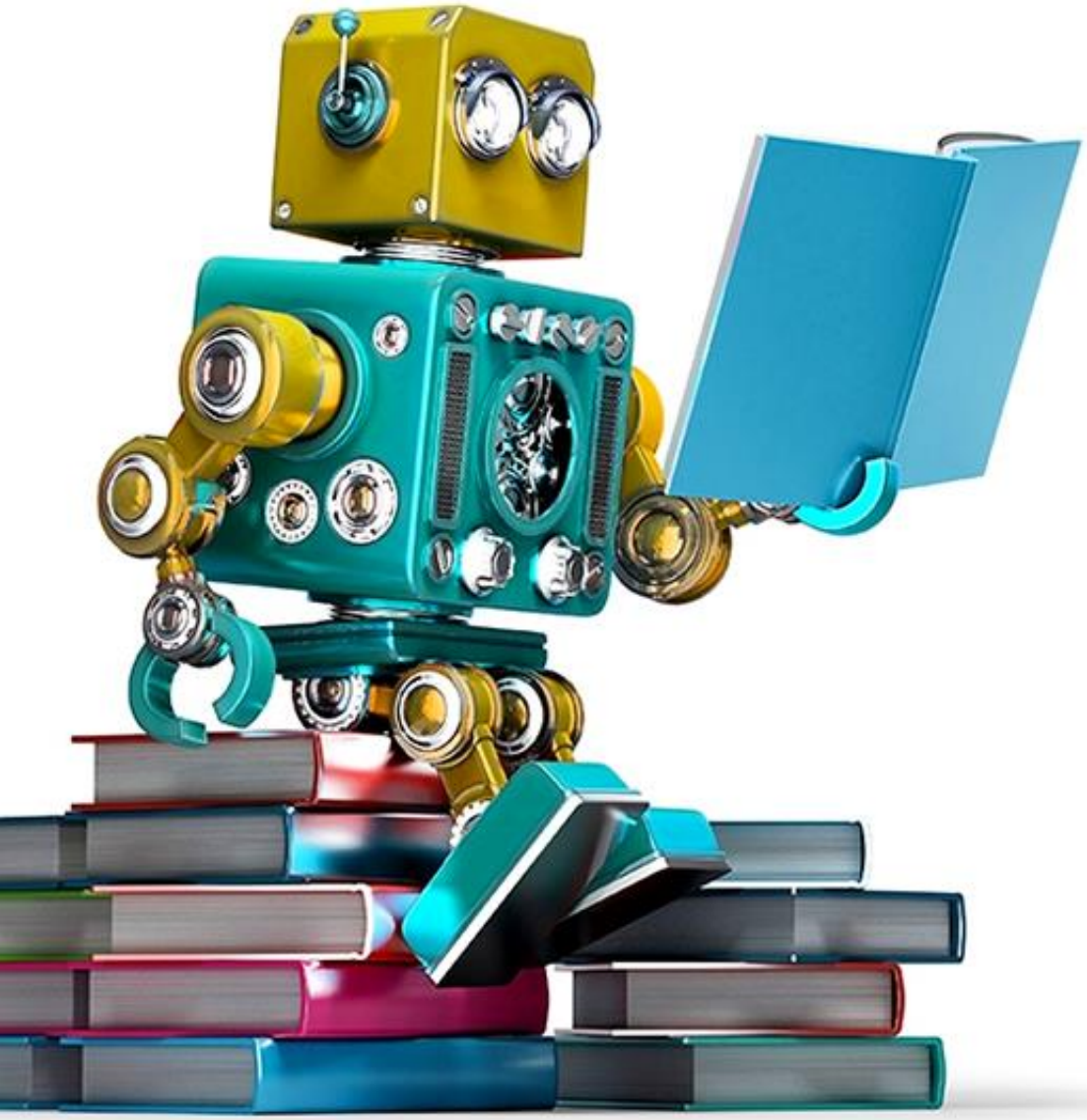


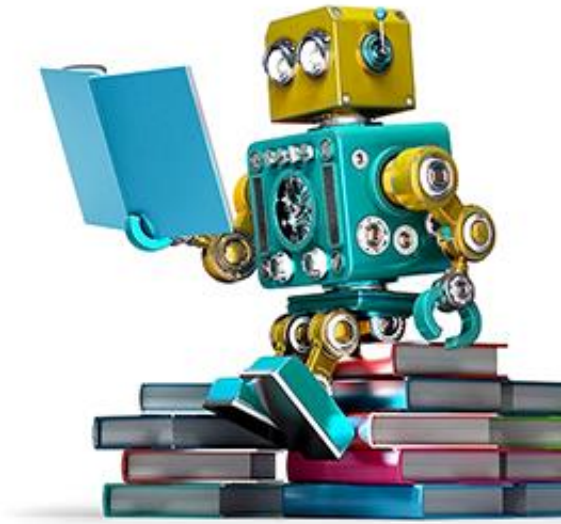
Anomaly Detection of Cellular network

Majid Hosseini Ph.D., P.Eng.

Research Engineer and Data Scientist

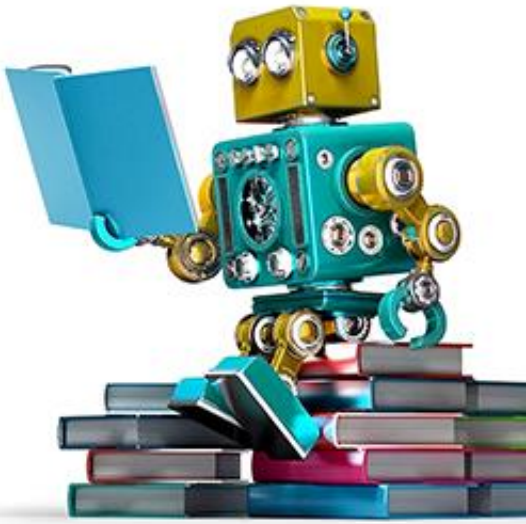
Objectives:

- Explore possibilities of ML to detect abnormal behaviors in the utilization of the network that would motivate a change in the configuration of the base station.
- Network optimization is aimed to train an ML system capable of classifying current activity as:
  - 0 (normal): normal behavior of any working day
  - 1 (unusual): deviation from typical activities due to strike, demonstration, sports event, etc. which should trigger a reconfiguration of the base station.

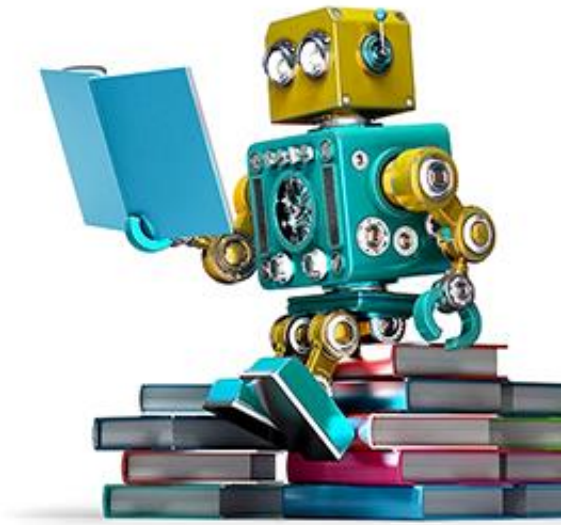


Anomaly Detection Methodology

- Exploratory Data Analysis was conducted to investigate:
  - distribution of unusual cases
  - Correlation Between Feature Variables
  - How to deal with non-quantitative features and missing values
- PCA and t-SNE Analysis conducted to investigate clustering
- Supervised Classification Modeling
  - XGBoost with optimized parameters provided excellent prediction
- Unsupervised Anomaly Detection
  - Multiple methods were implemented and optimized for improved performance



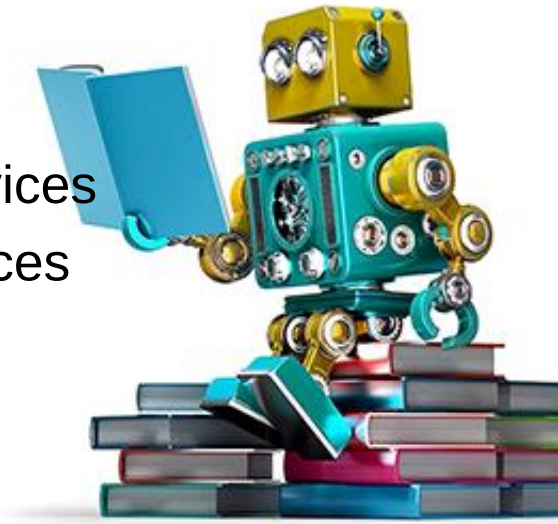
Exploratory Data Analysis



Dataset

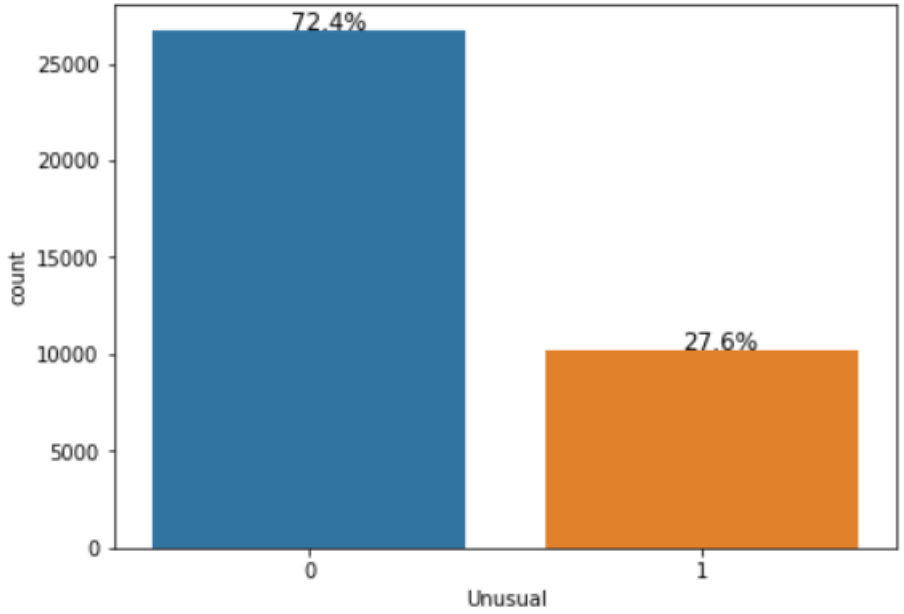
- Dataset has been obtained from a real LTE deployment
- Following metrics were gathered from a set of 10 base stations, every 15 minutes, for two weeks
  - **Time** : hour of the day when the sample was generated.
  - **CellName**: text string used to uniquely identify the cell
  - **PRBUsageUL** and **PRBUsageDL**: level of resource utilization
  - **meanThrDL** and **meanThrUL**: average carried traffic (in Mbps)
  - **maxThrDL** and **maxThrUL**: maximum carried traffic (in Mbps)
  - **meanUEDL** and **meanUEUL**: average no. of user equipment (UE) devices
  - **maxUEDL** and **maxUEUL**: maximum no. of user equipment (UE) devices
  - **Unusual**: labels for supervised learning.

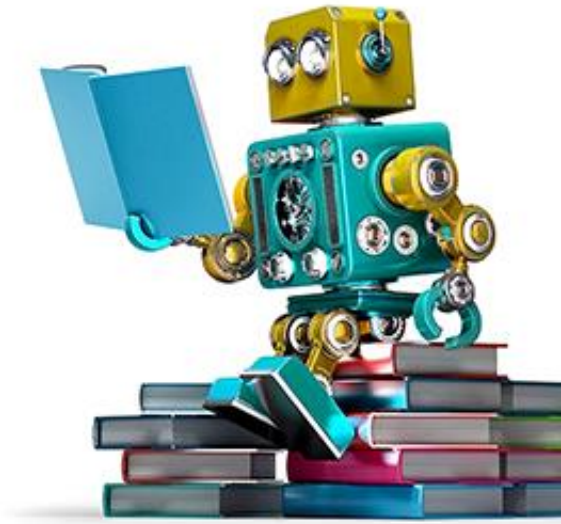
UL: Uplink and DL: downlink



Dataset

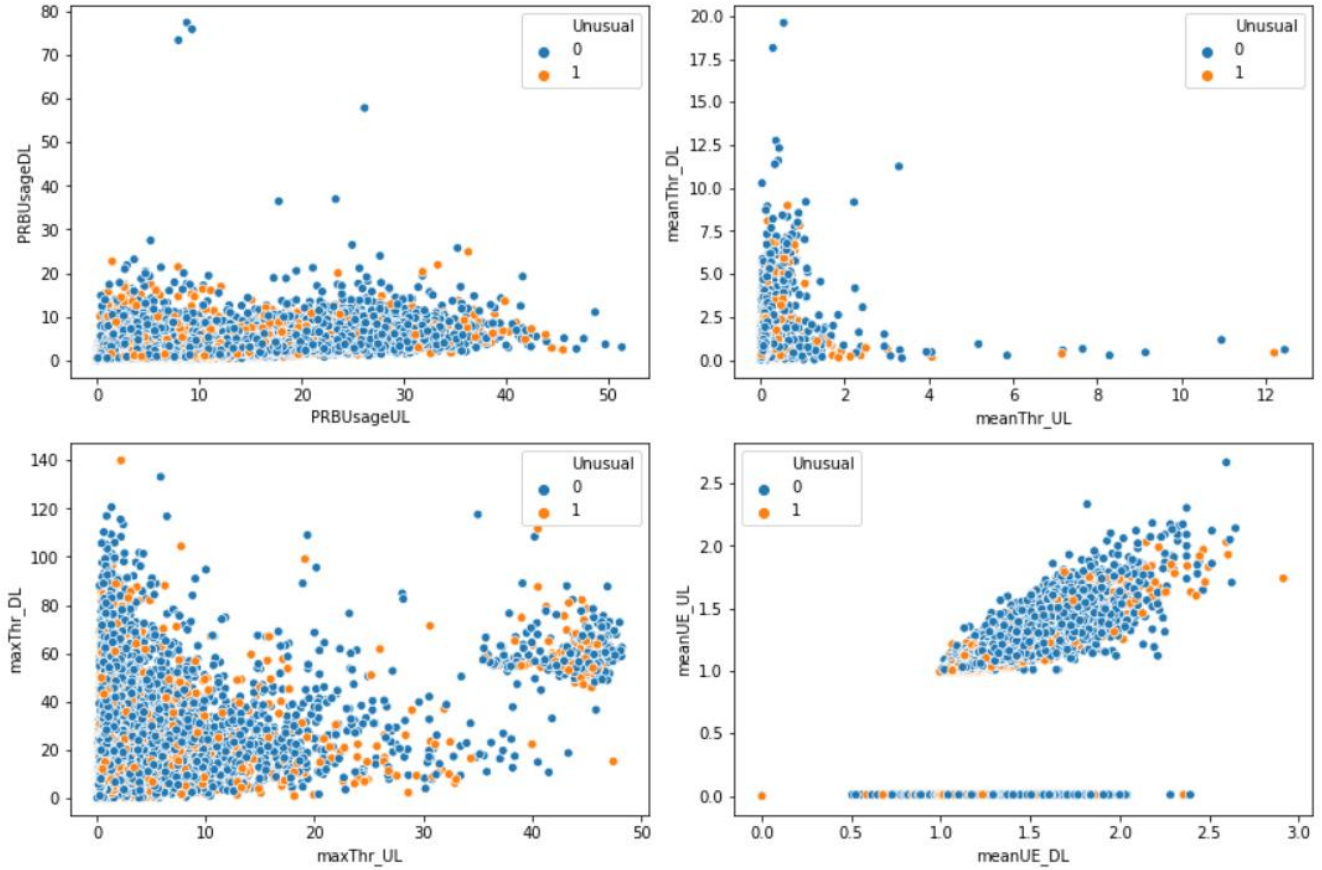
- Dataset composed of 27.6% unusual cases (anomalies) as indicated in following count plot

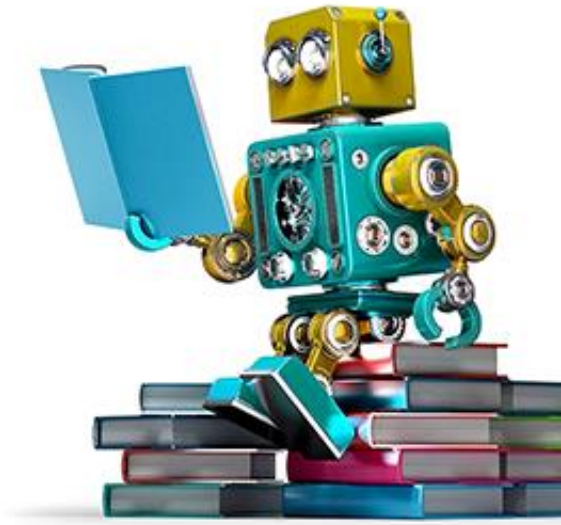




Visual Representation of Corresponding Features

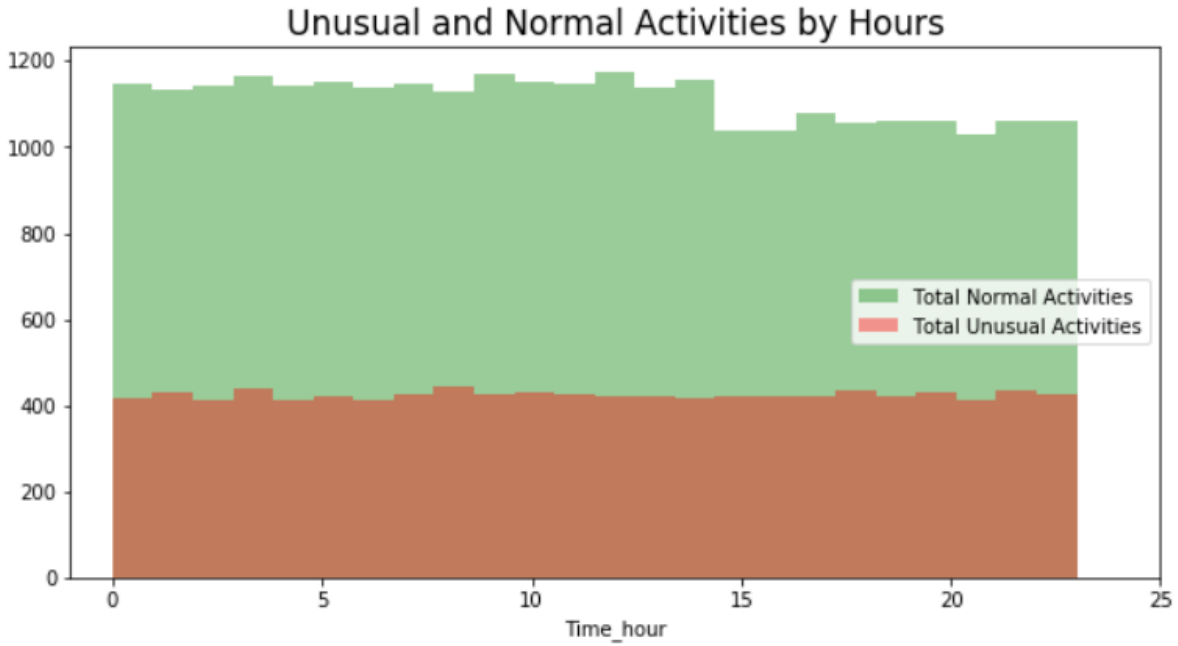
- The categorization of features based on "Normal" and "Unusual“ do not indicate any obvious separation boundary



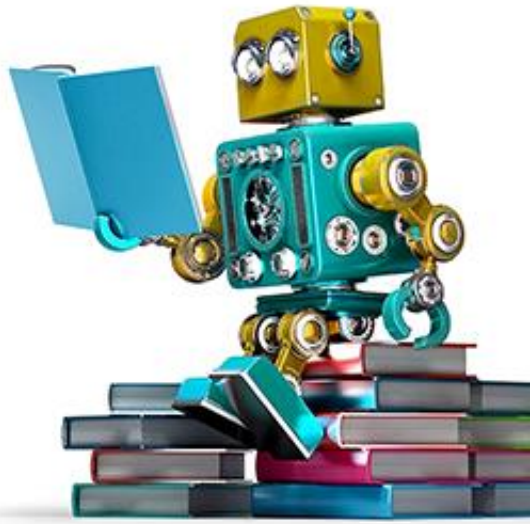


Investigation of “Time” Feature

- Hour of the day when the sample was generated was extracted from "Time" feature and plotted for both Normal and Unusual activities

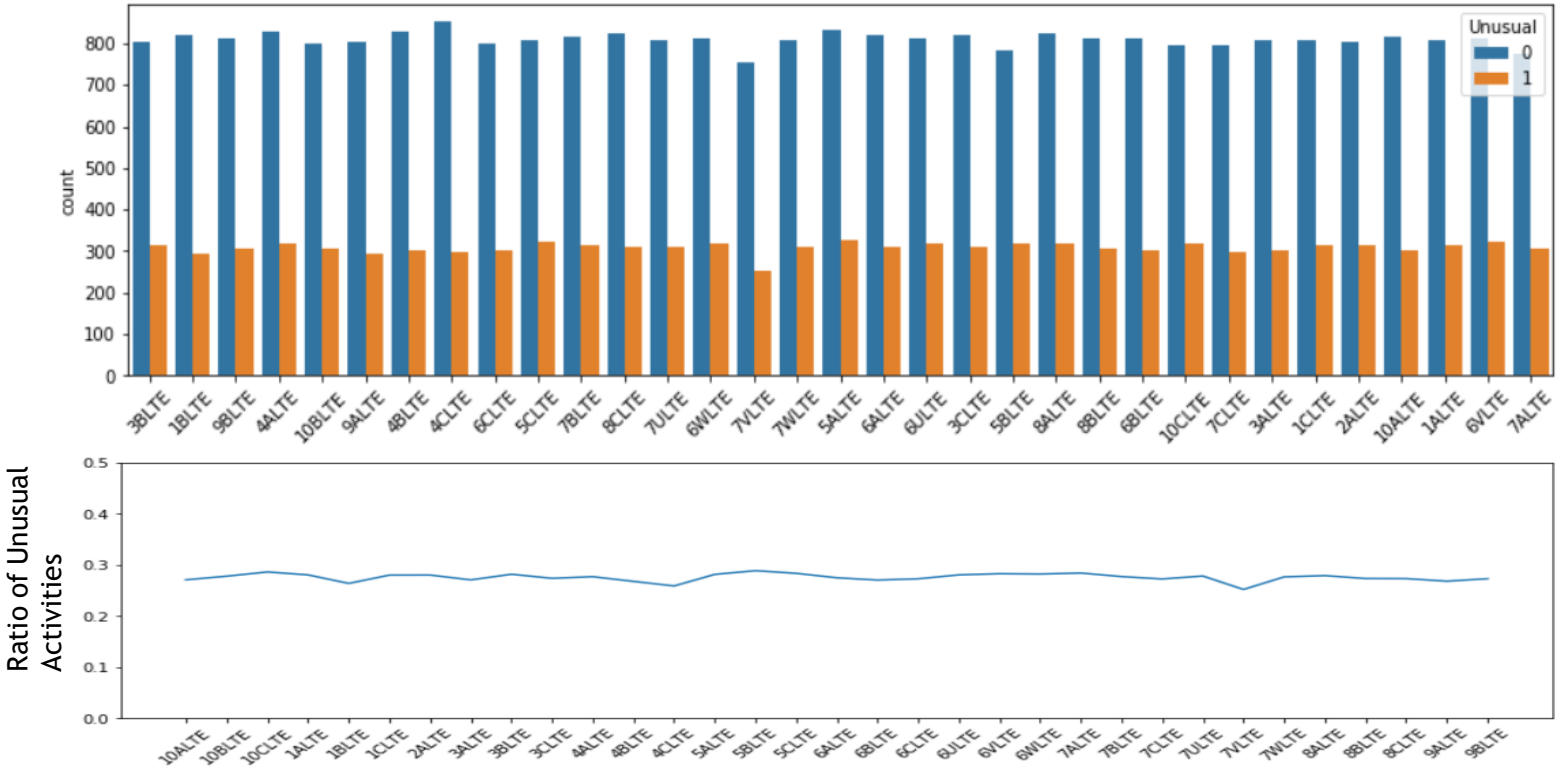


No clear relationship is evident from the plot and in fact, the total activities are extremely similar across hours. So, the time column was dropped from dataset

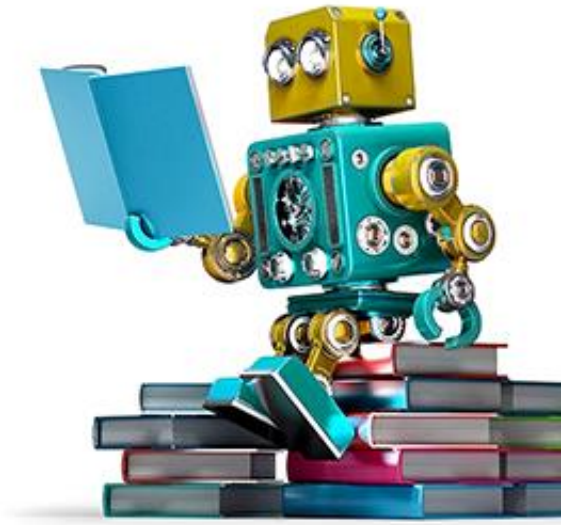


CellName Relevance Investigation

- Relationship of feature "CellName" with Normal and Unusual activities is plotted in following



No clear relationship is evident from the plot and the "CellName" column was dropped from dataset

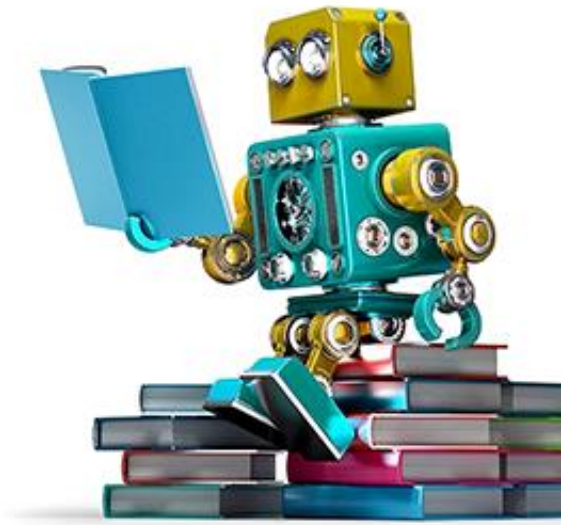


Missing Data

- Following is a list of missing values and their column names

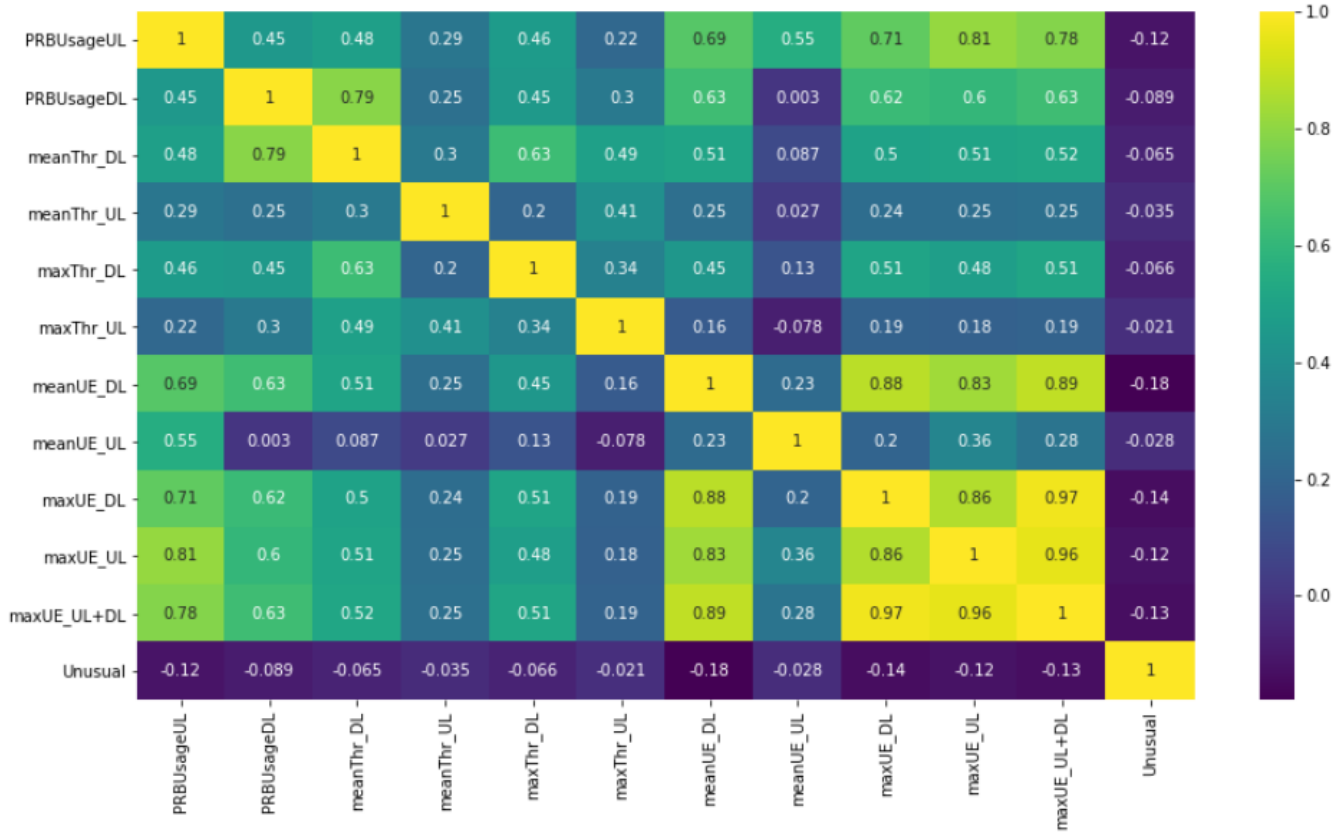
	Total Missings	Missing Count %	Total Missings (Unusual)	Missing Count (Unusual) %
maxUE_UL	89	0.2	16	0.87
maxUE_DL	89	0.2	16	0.87

- There are ~0.2% data recording with missing values, among them there are 16 Unusual cases which is equivalent if less than 1% of Unusual cases.
- Considering that 27.6% of total dataset are Unusual cases, it makes sense to drop the rows with missing values.

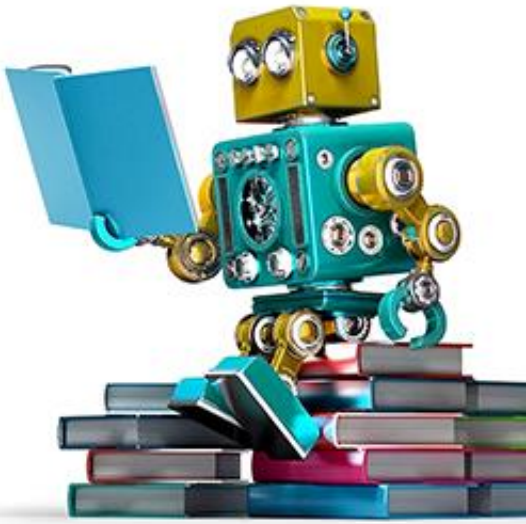


Correlation Between Feature Variables

- Heatmap of correlations:



Based on the heatmap the features “maxUEDL”, “maxUEUL”, and “maxUE_UL+DL” are highly correlated.



Multicollinearity

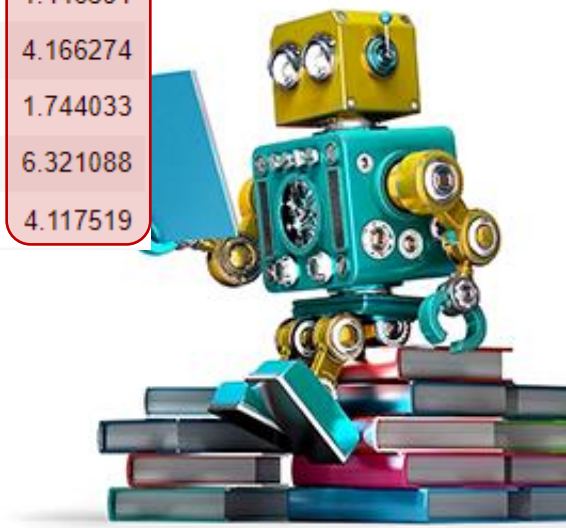
- The highly correlated features is an indication of redundancy of features and pose a potential risk to the accuracy of ML models
- Variation Inflation Factor (VIF) utilized to quantify the degree of Multicollinearity

	Feature	VIF
0	PRBUsageUL	7.682292
1	PRBUsageDL	7.291296
2	meanThr_DL	7.177465
3	meanThr_UL	1.447954
4	maxThr_DL	4.396230
5	maxThr_UL	1.746526
6	meanUE_DL	27.535386
7	meanUE_UL	4.482116
8	maxUE_DL	inf
9	maxUE_UL	inf
10	maxUE_UL+DL	inf

	Feature	VIF
0	PRBUsageUL	7.682292
1	PRBUsageDL	7.291296
2	meanThr_DL	7.177465
3	meanThr_UL	1.447954
4	maxThr_DL	4.396230
5	maxThr_UL	1.746526
6	meanUE_DL	27.535386
7	meanUE_UL	4.482116
8	maxUE_UL	82.292923
9	maxUE_UL+DL	113.382678

	Feature	VIF
0	PRBUsageUL	7.219581
1	PRBUsageDL	7.052079
2	meanThr_DL	7.040490
3	meanThr_UL	1.447886
4	maxThr_DL	4.229949
5	maxThr_UL	1.744822
6	meanUE_DL	18.668878
7	meanUE_UL	4.124259
8	maxUE_UL	30.404953

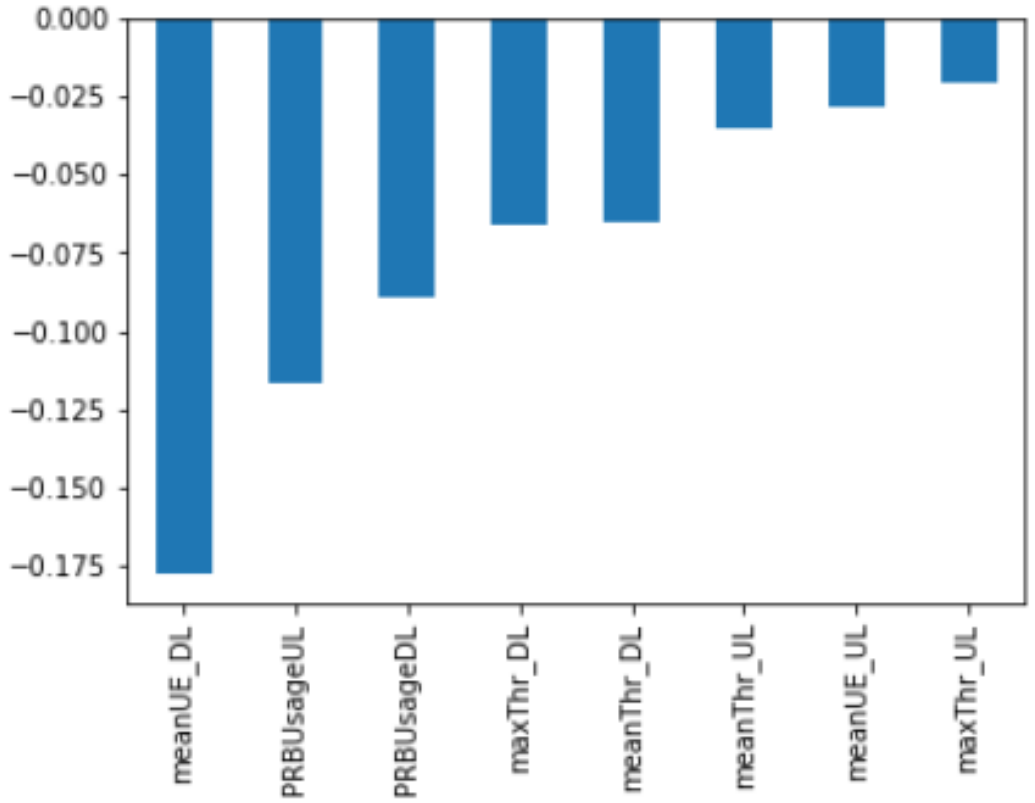
	Feature	VIF
0	PRBUsageUL	4.231607
1	PRBUsageDL	6.356108
2	meanThr_DL	6.850264
3	meanThr_UL	1.446591
4	maxThr_DL	4.166274
5	maxThr_UL	1.744033
6	meanUE_DL	6.321088
7	meanUE_UL	4.117519

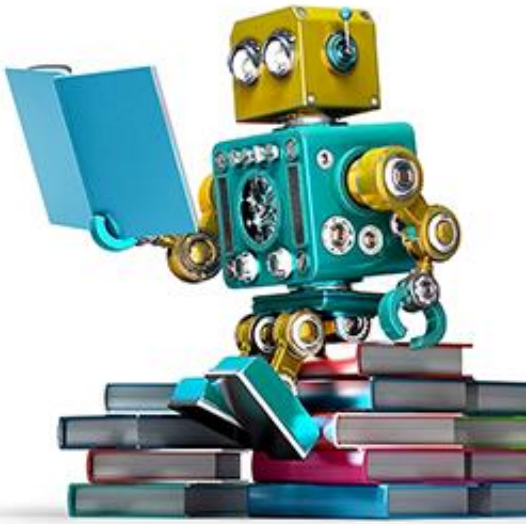


inf for VIF indicates a perfect correlation, thus columns can be dropped.
Ideally, VIF should be < 10

Correlation of Feature with Target Variables

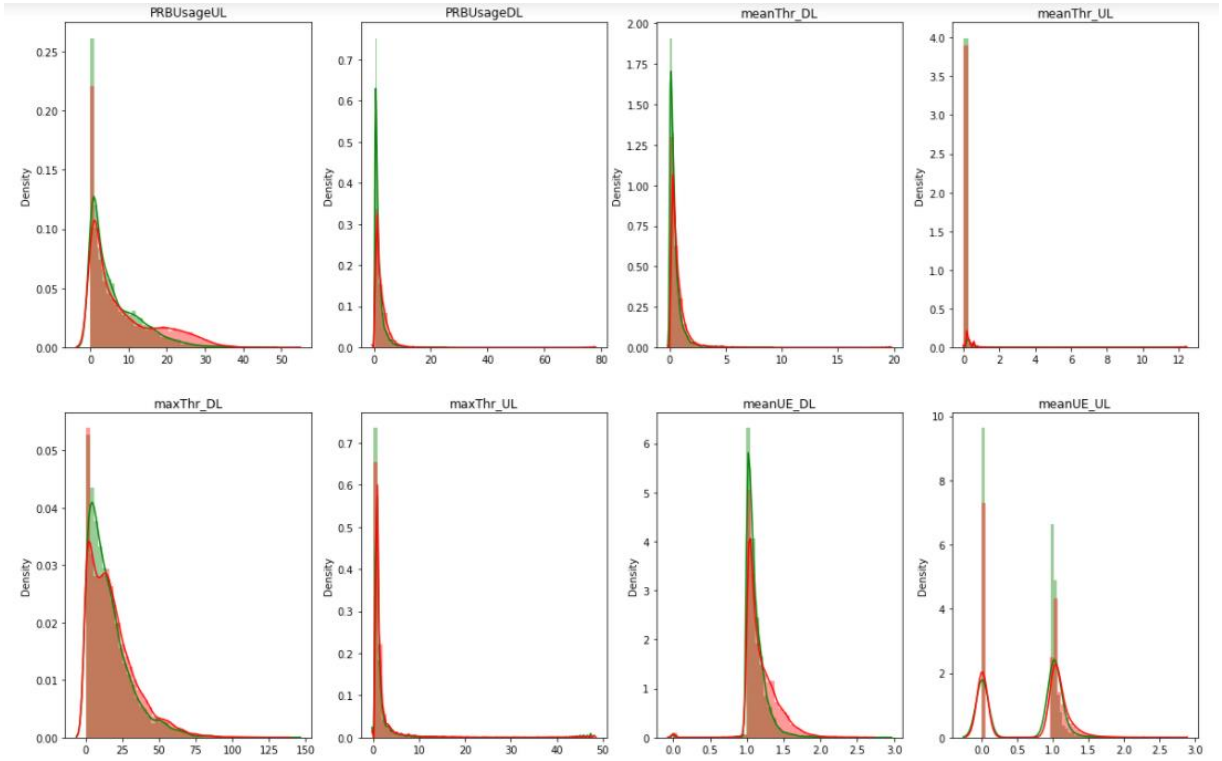
- Bar plot of correlations:
  - Should the features with low dependency be dropped from dataset?

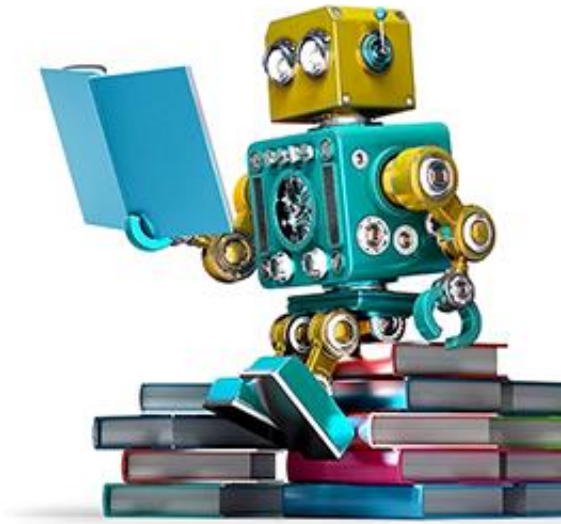




Feature Selection using Z-test

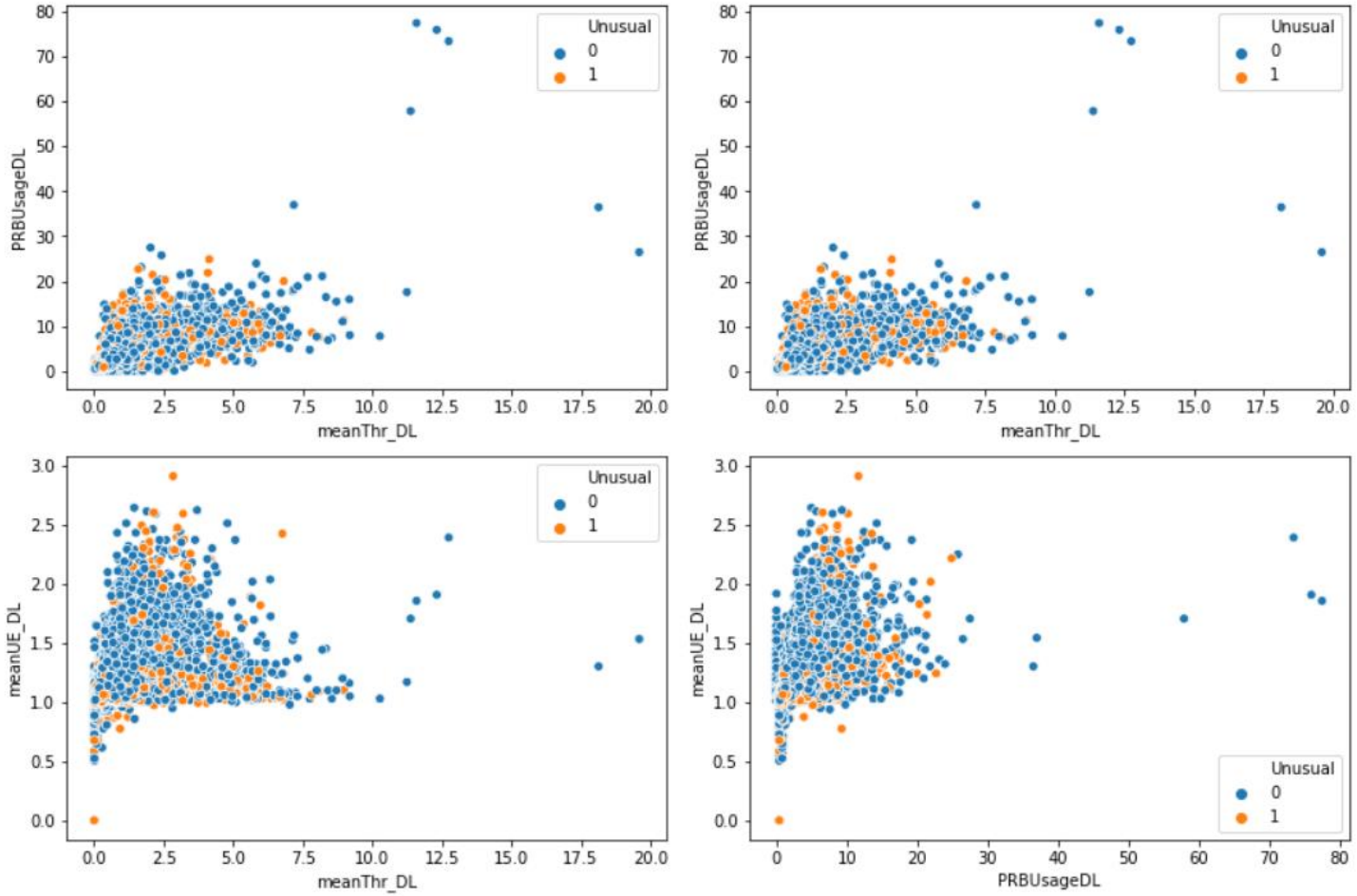
- Hypothesis testing indicated that all features are statistically significant
  - Tested with significance level of 0.05 (two tailed test)
- Thus, all features kept in the model

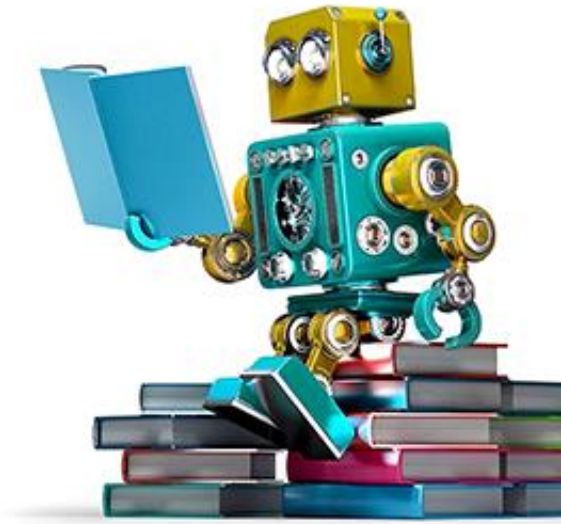




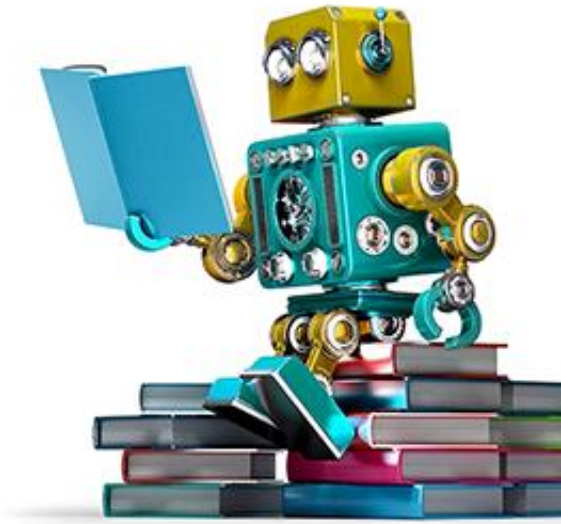
Relationship of Important Features

- The categorization of features based on "Normal" and "Unusual“ do not indicate any obvious separation boundary



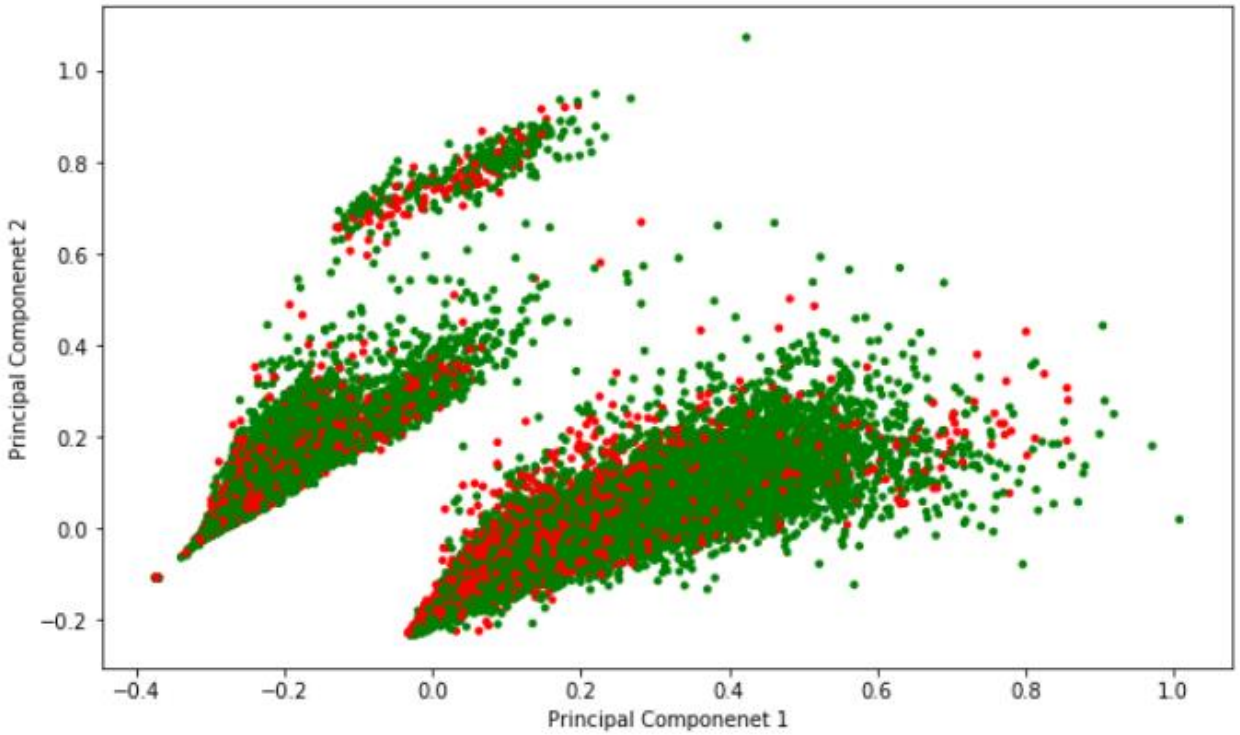


PCA and t-SNE Analysis

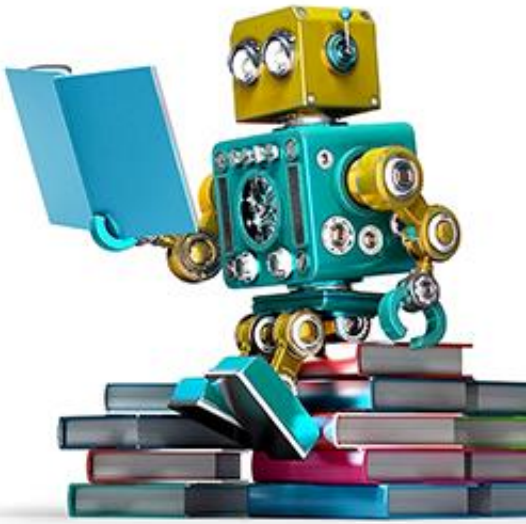


2D PCA Plot

- Features were scaled using MinMaxScaler() before applying PCA decomposition
- PCA tries to provide the projection using the correlation between some dimensions and keeping the maximum amount of information about the original data distribution.

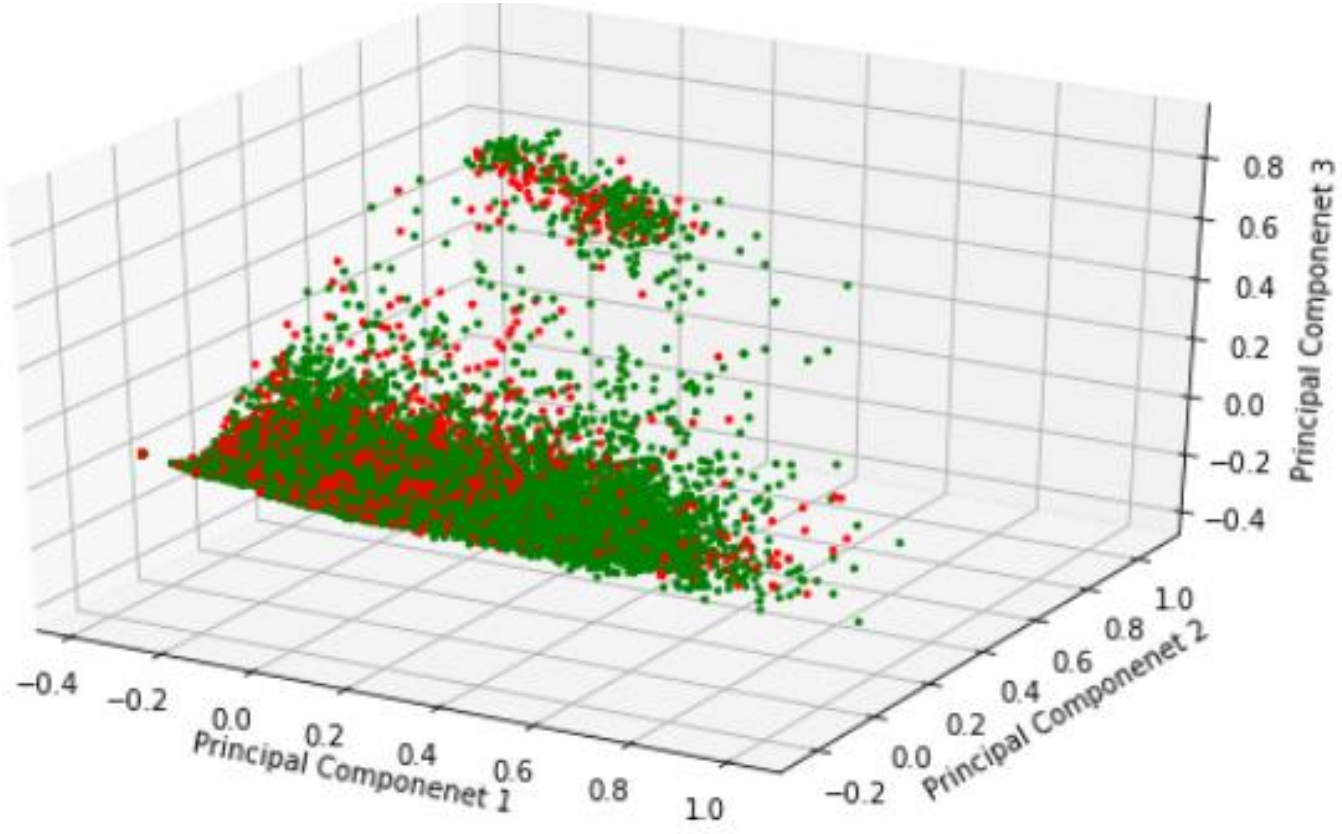


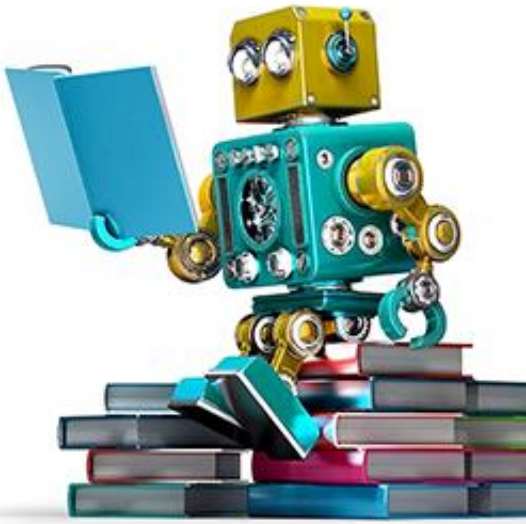
There are not clear clusters nor a clear defined pattern.



3D PCA Plot

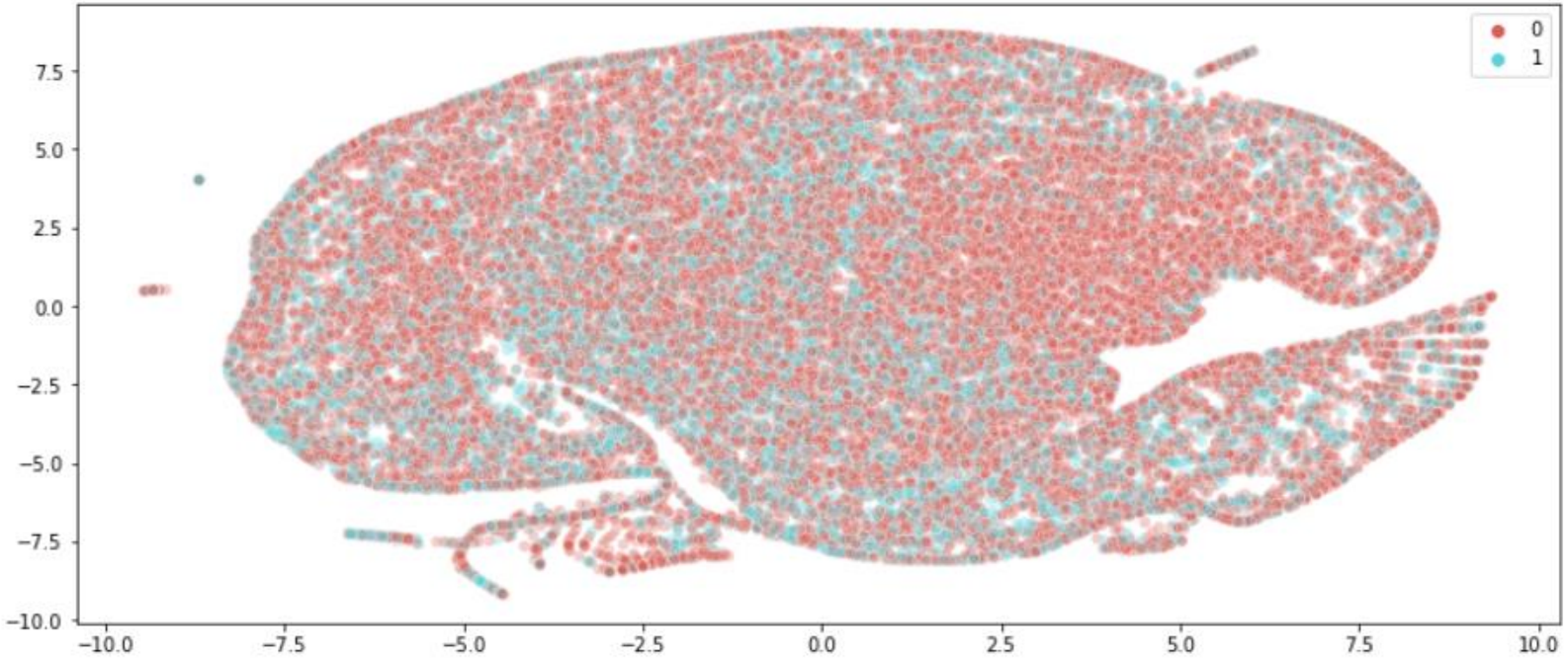
- Normal and unusual samples are mixed in many with no clear decision boundaries
- It limit the options for the classifier that is going to be used, e.g linear classifiers can be directly excluded.



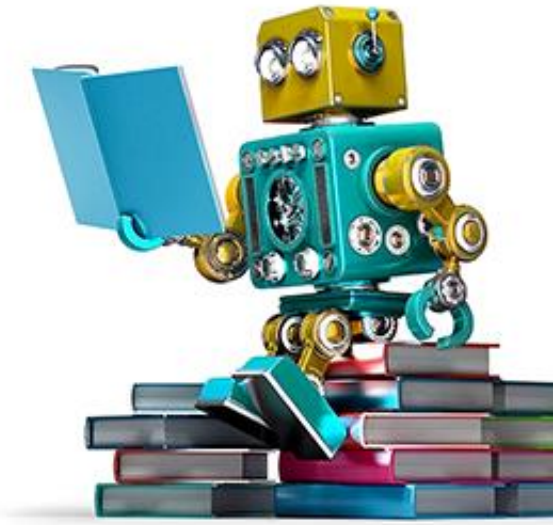


t-SNE Plot

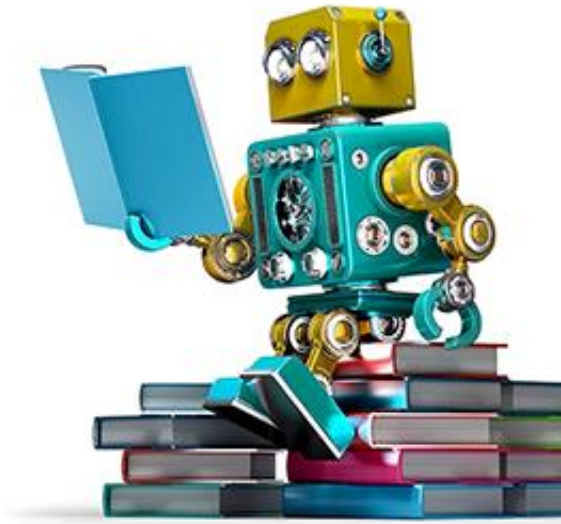
- t-SNE is an unsupervised non-linear reduction method to minimize the divergence between a distribution that measures pairwise similarities of the input and a distribution that measures the similarities of the corresponding low-dimensional points in the embedding.



t-SNE has built a set of separable clusters but with samples of different classes mixed in the same clusters, without a clear visual pattern.

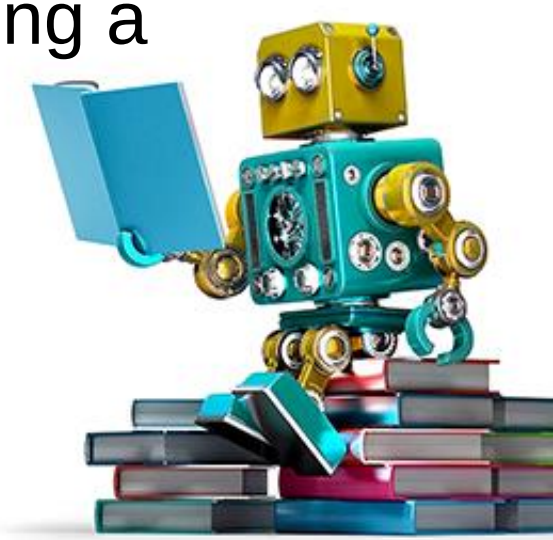


Supervised Classification Modeling



eXtreme Gradient Boosting (XGBoost)

- XGBoost algorithm indicated a superior performance for classification tasks as demonstrated in my previous project (available on GitHub):
  - Comprehensive Evaluation of Machine Learning Techniques for Supervised Classification Tasks
- Train/Test split was performed test_size=0.2 and forcing a similar distribution of Unusual cases (stratify=y).
- XGBoost has been implemented and the parameters optimized using a cross-validated grid-search over a parameter grid

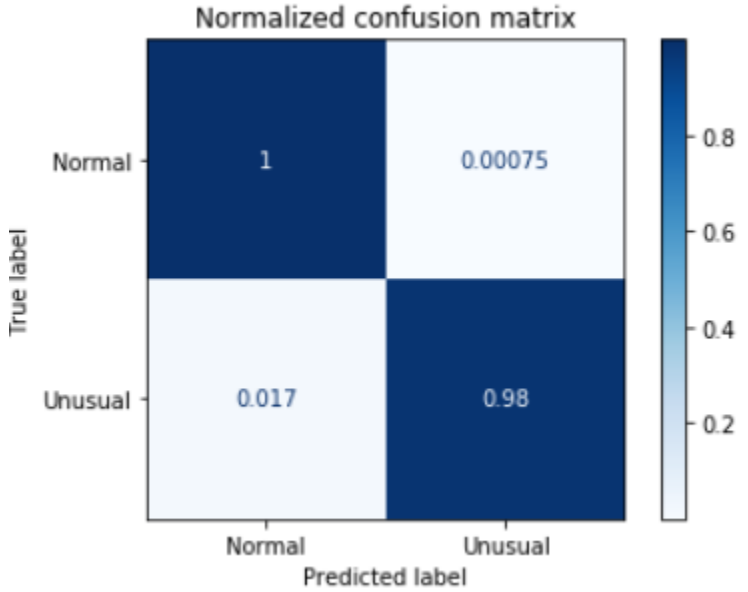


eXtreme Gradient Boosting (XGBoost)

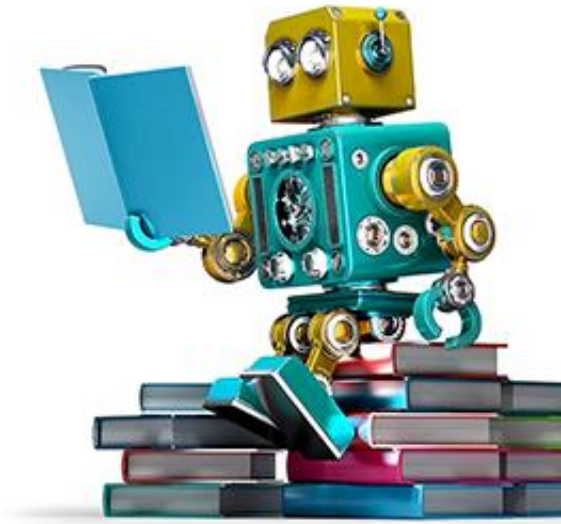
- Confusion matrix of optimized XGBoost for highest recall_score
- XGBoost demonstrated *exceptional prediction power* to classify *unusual activities*

	Predcited: Normal	Predcited: Unusual
Actual: Normal	TN = 5326	FP = 4
Actual: Unusual	FN = 34	TP = 1999

	precision	recall	f1-score	support
Normal	0.993657	0.999250	0.996445	5330.0
Unusual	0.998003	0.983276	0.990585	2033.0
avg/Total	0.994857	0.994839	0.994827	7363.0



There are 34 FN that can be reduced by modifying of threshold

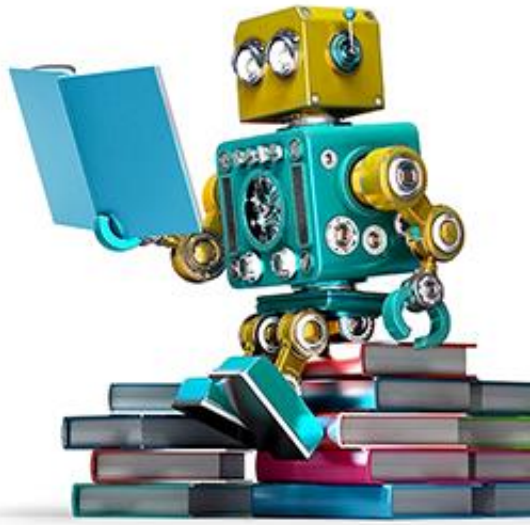


Threshold Modifications

- Classifier Threshold can be adjusted for a trade-off between FN and FP
- Typically, in anomaly detection FN is rather an unacceptable false than FP.
  - classifier allows malicious traffic to network system or accept a fraud credit card activity
- Here, the FN decreased from 34 to 13 cases by adjusting threshold from 0.5 to 0.1

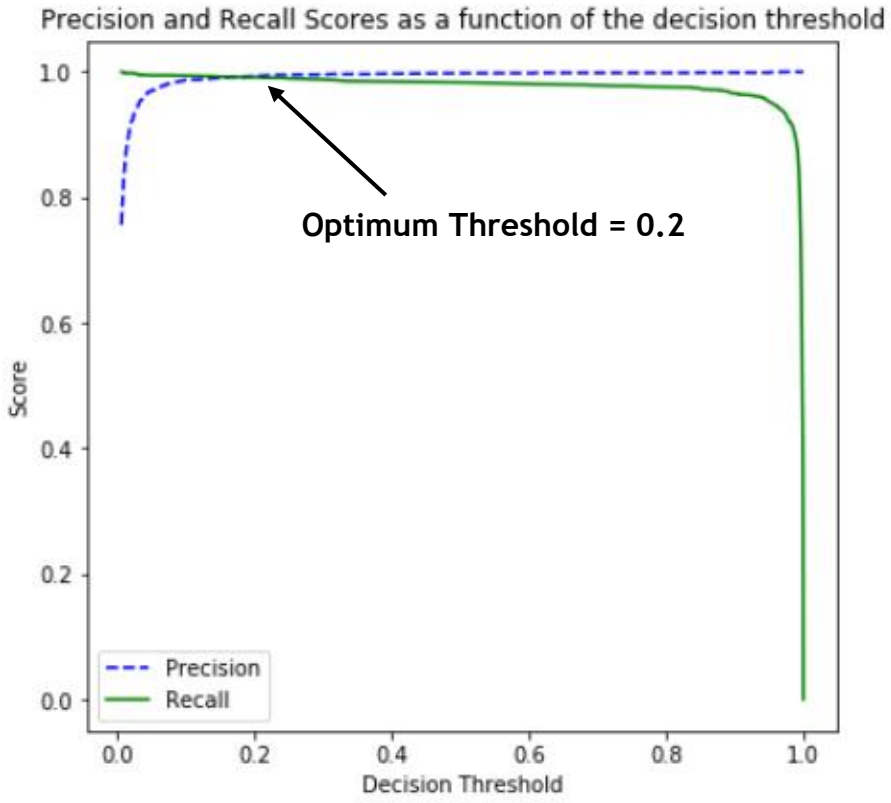
	Predcited: Normal	Predcited: Unusual
Actual: Normal	TN = 5303	FP = 27
Actual: Unusual	FN = 13	TP = 2020

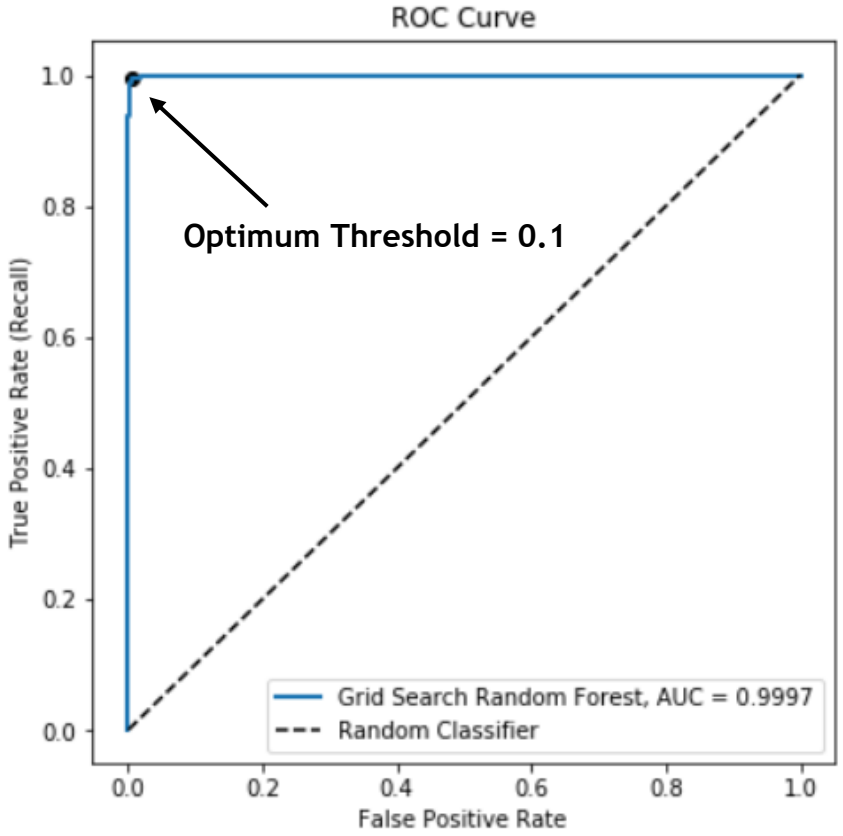
	precision	recall	f1-score	support
Normal	0.997555	0.994934	0.996243	5330.0
Unusual	0.986810	0.993606	0.990196	2033.0
avg/Total	0.994588	0.994567	0.994573	7363.0

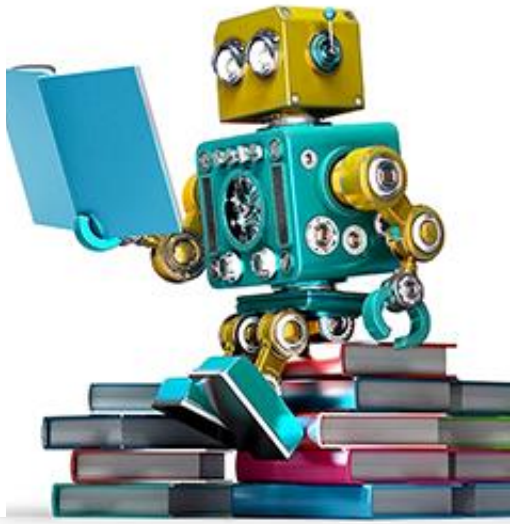


Threshold Tuning

- Optimal threshold can be identified based on precision-recall curve (highest F1-score) or ROC curve (plot of hit rate vs false alarm rate)

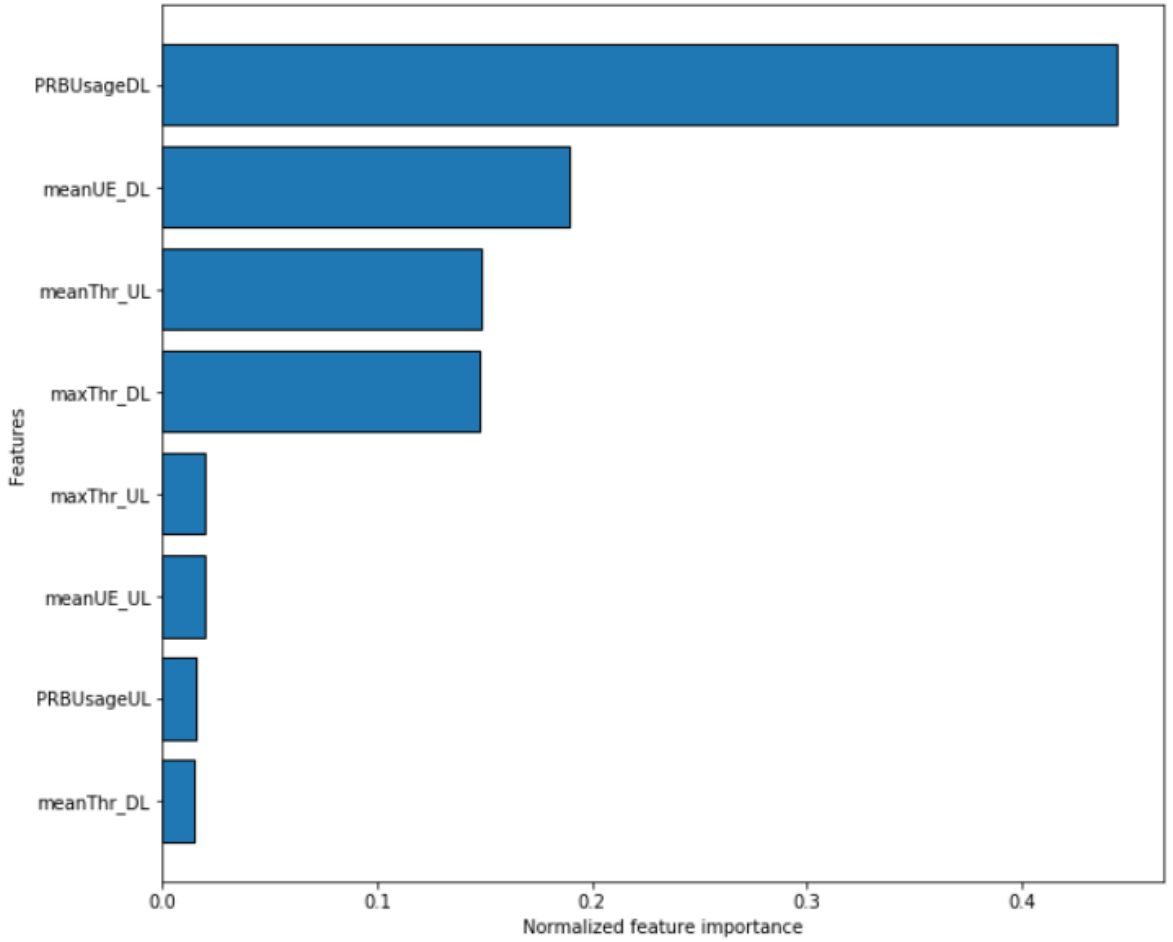


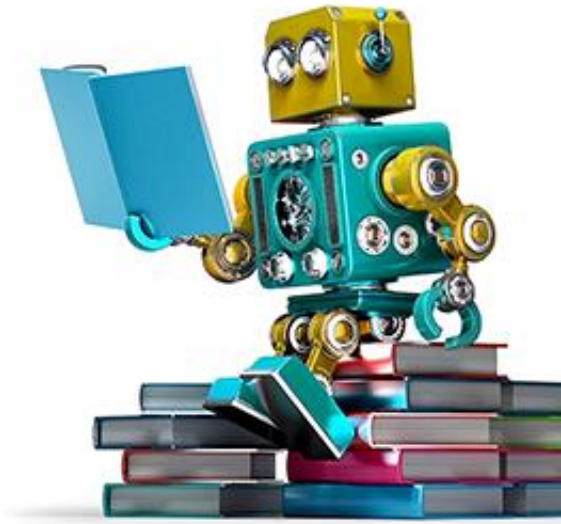




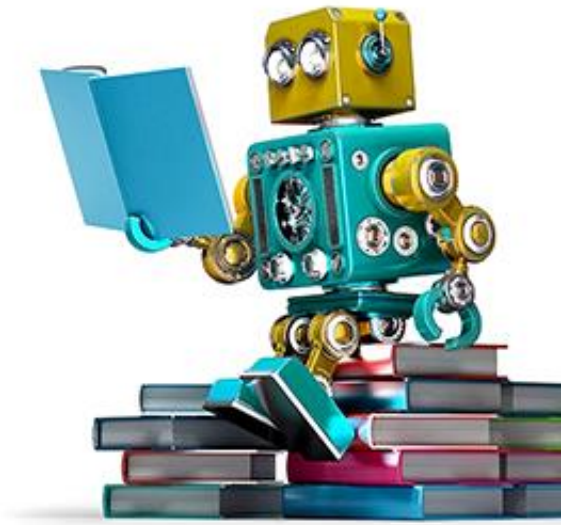
Feature Importance

- The normalized importance of different features identified by the trained XGBoost algorithm is compared as a bar plotted



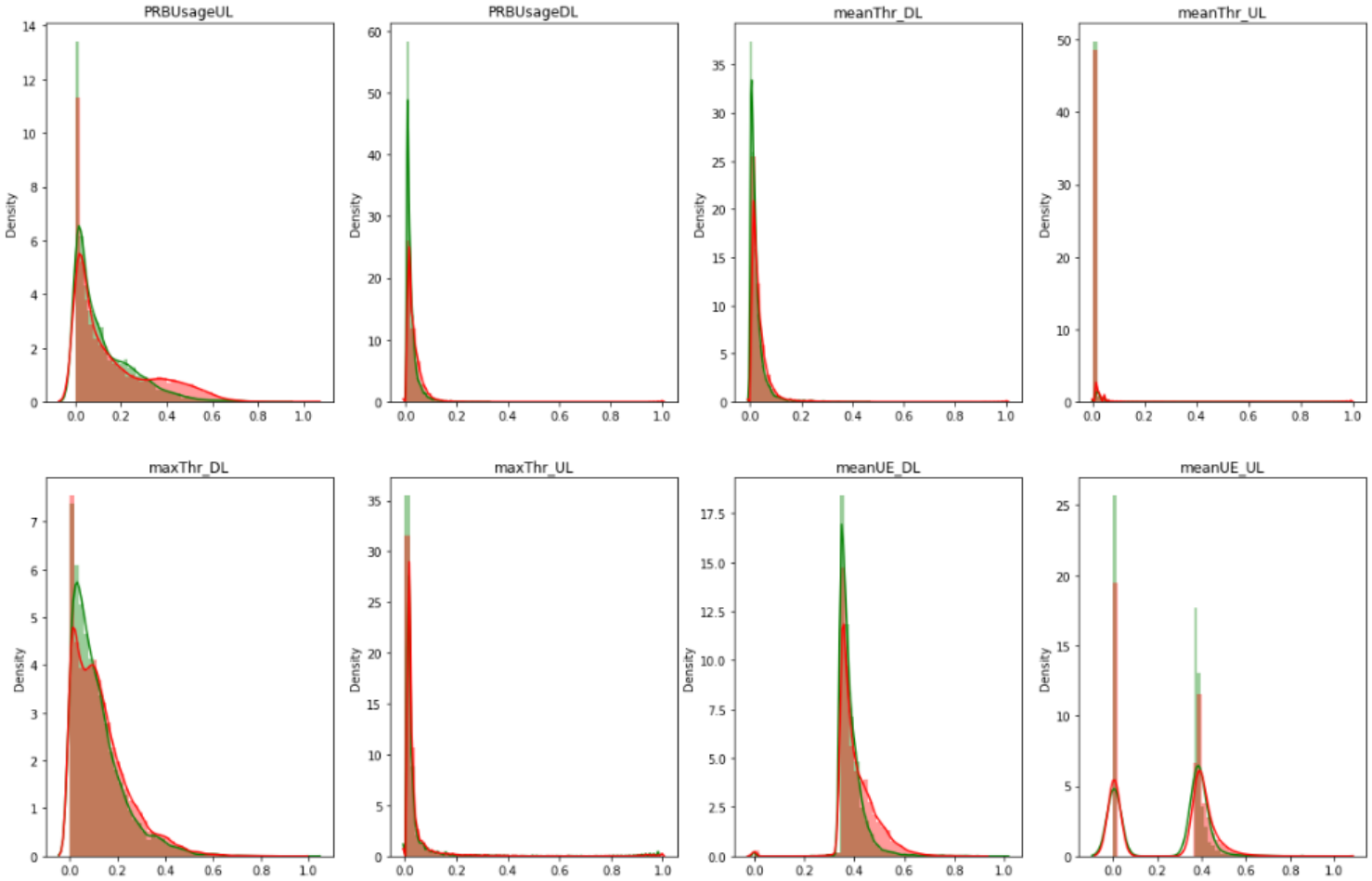


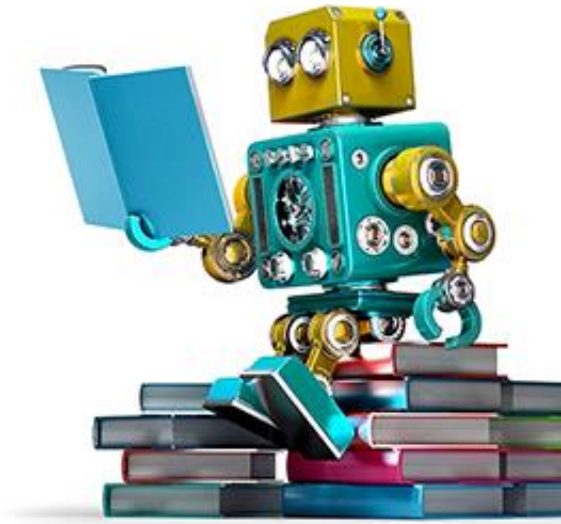
Unsupervised Anomaly Detection



Distribution of Normalized Features

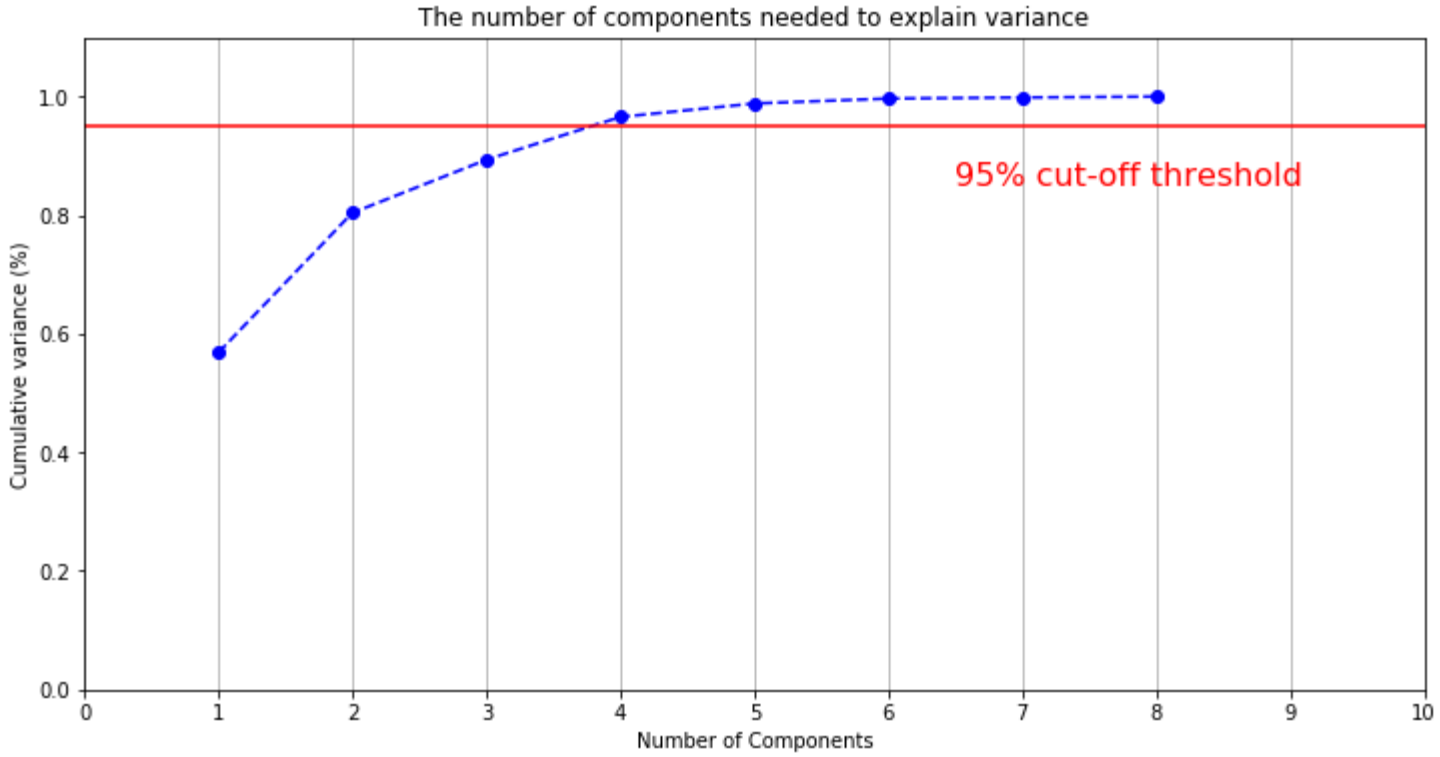
- Substantial overlap of unusual and normal activities

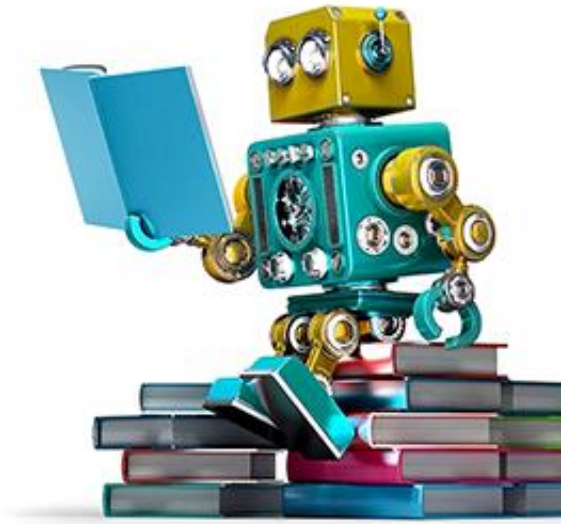




PCAs with 95% Explained Variance

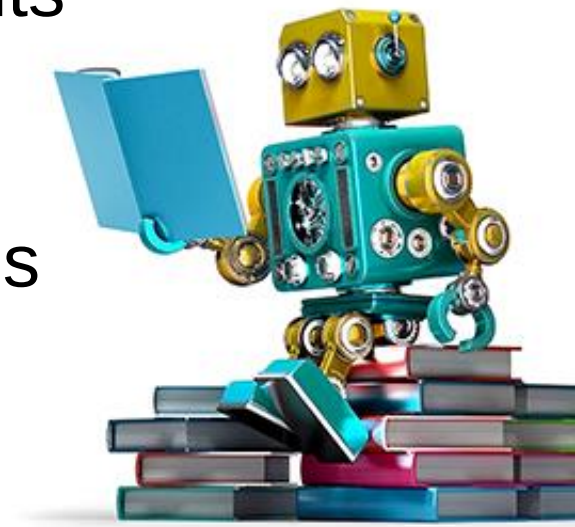
- A PCA decomposition with 4 components can explain 95% of variance of data

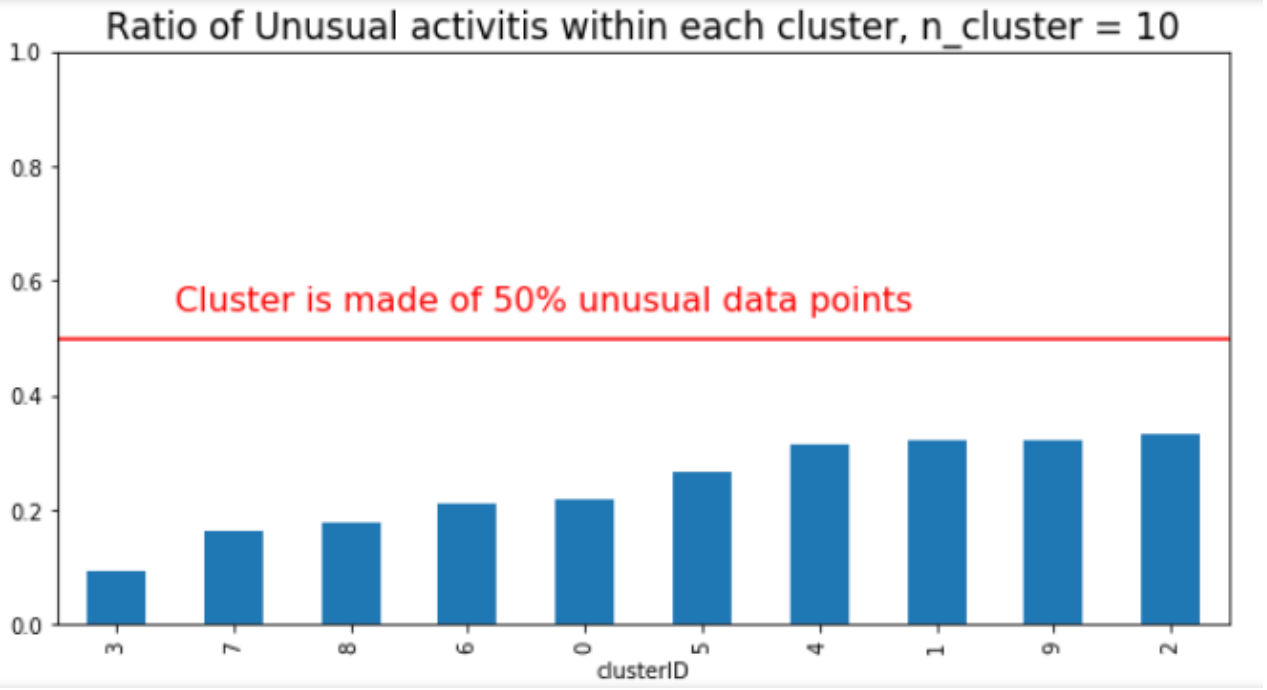


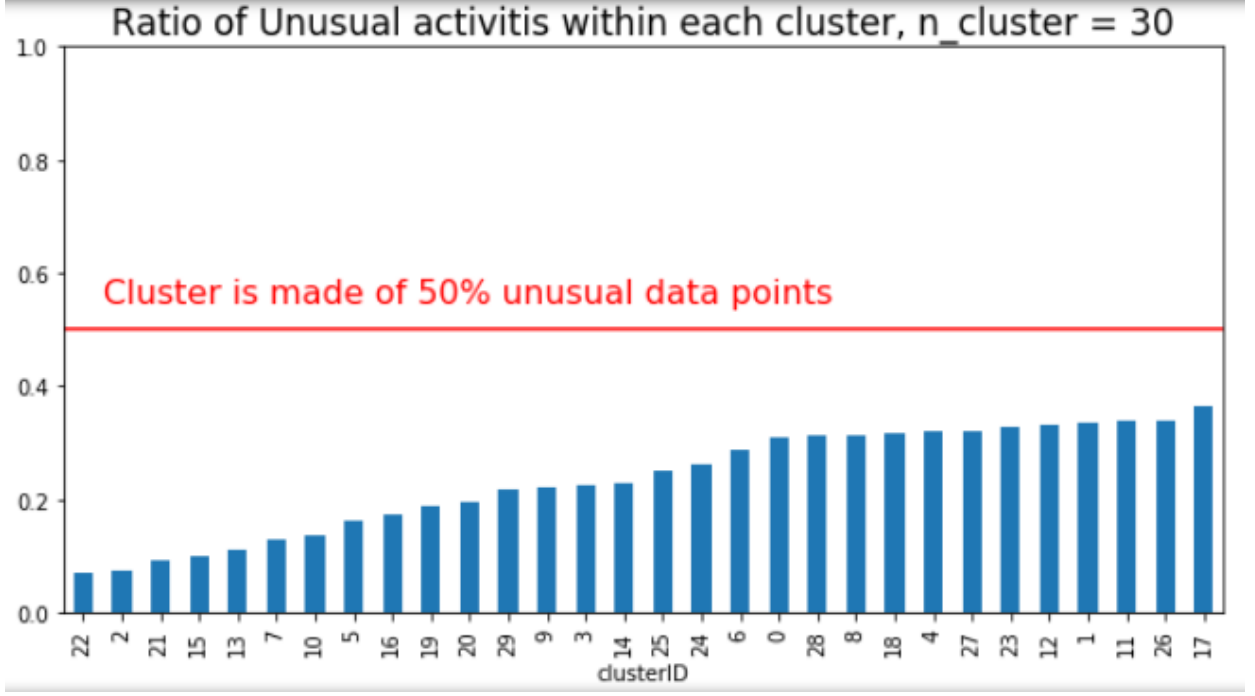


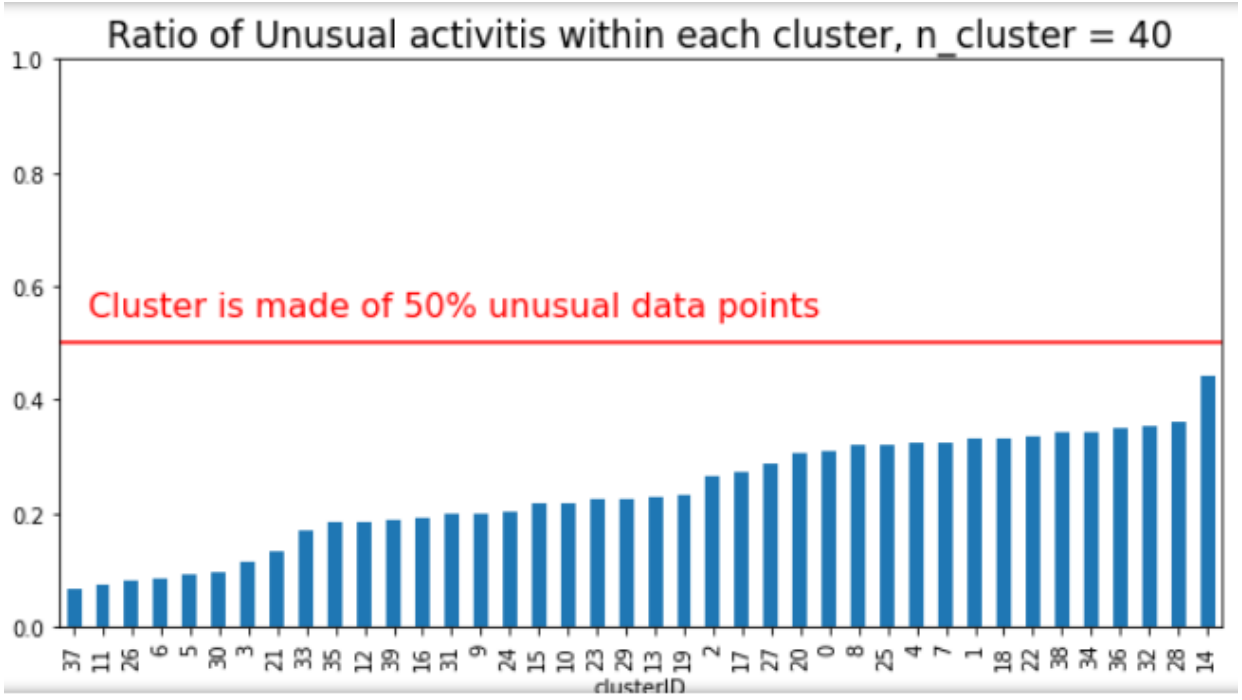
K-means Clustering

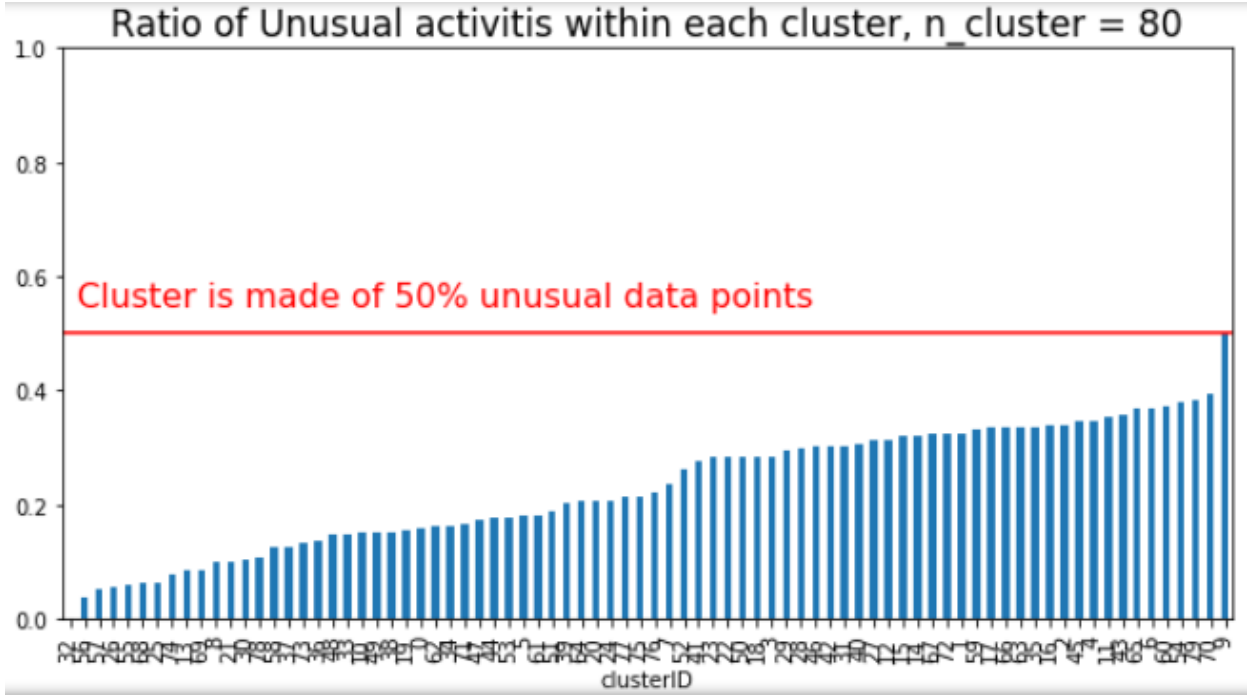
- The KMeans algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares
- First task was to investigate if we were able to build a cluster mainly composed of unusual data points
- Ratio of the unusual data points to the total data points within each cluster was plotted for following range
  - n_cluster = 5 to n_cluster = 100
- A red line added to each plot indicating if the cluster is made of 50% unusual data point




Ratio of Unusual activitis within each cluster, n_cluster = 10
Cluster is made of 50% unusual data points
1.0
0.8
0.6
0.4
0.2
0.0
3
7
8
6
0
5
4
1
9
2
clusterID

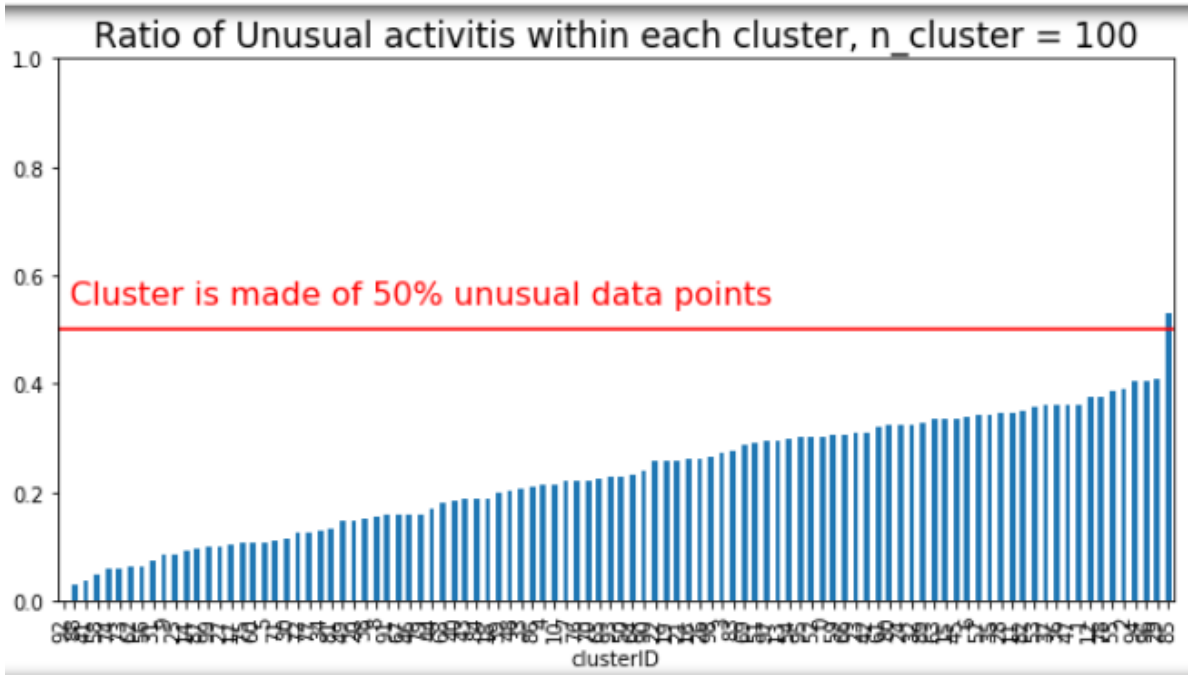

Ratio of Unusual activitis within each cluster, n_cluster = 30
Cluster is made of 50% unusual data points
1.0
0.8
0.6
0.4
0.2
0.0
22
2
21
15
13
7
10
5
16
19
20
29
9
3
14
25
24
6
0
28
8
18
4
27
23
12
1
11
26
17
clusterID


Ratio of Unusual activitis within each cluster, n_cluster = 40
Cluster is made of 50% unusual data points
1.0
0.8
0.6
0.4
0.2
0.0
37
11
26
6
5
30
3
21
33
35
12
39
16
31
9
24
15
10
23
29
13
19
2
17
27
20
0
8
25
4
7
1
18
22
38
34
36
32
28
14
clusterID

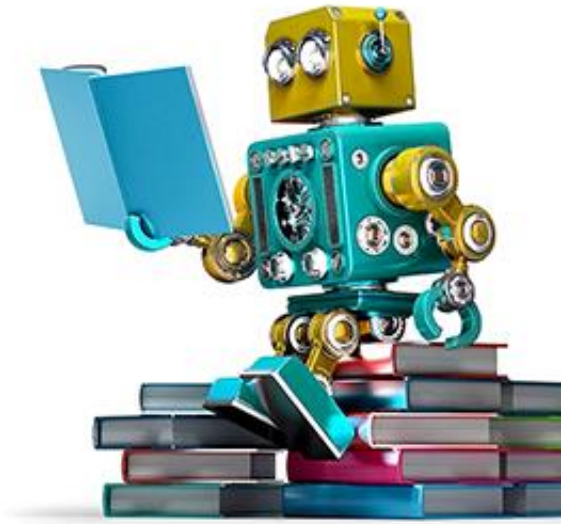

Ratio of Unusual activitis within each cluster, n_cluster = 80
Cluster is made of 50% unusual data points
1.0
0.8
0.6
0.4
0.2
0.0
clusterID

K-means Clustering

- The *prevalence* of unusual points in each cluster can be interpreted as *possibility* of an unseen point being *unusual* if assigned to this cluster
- Comparable results were achieved using both original scaled features and reduced features (4 components of PCA)

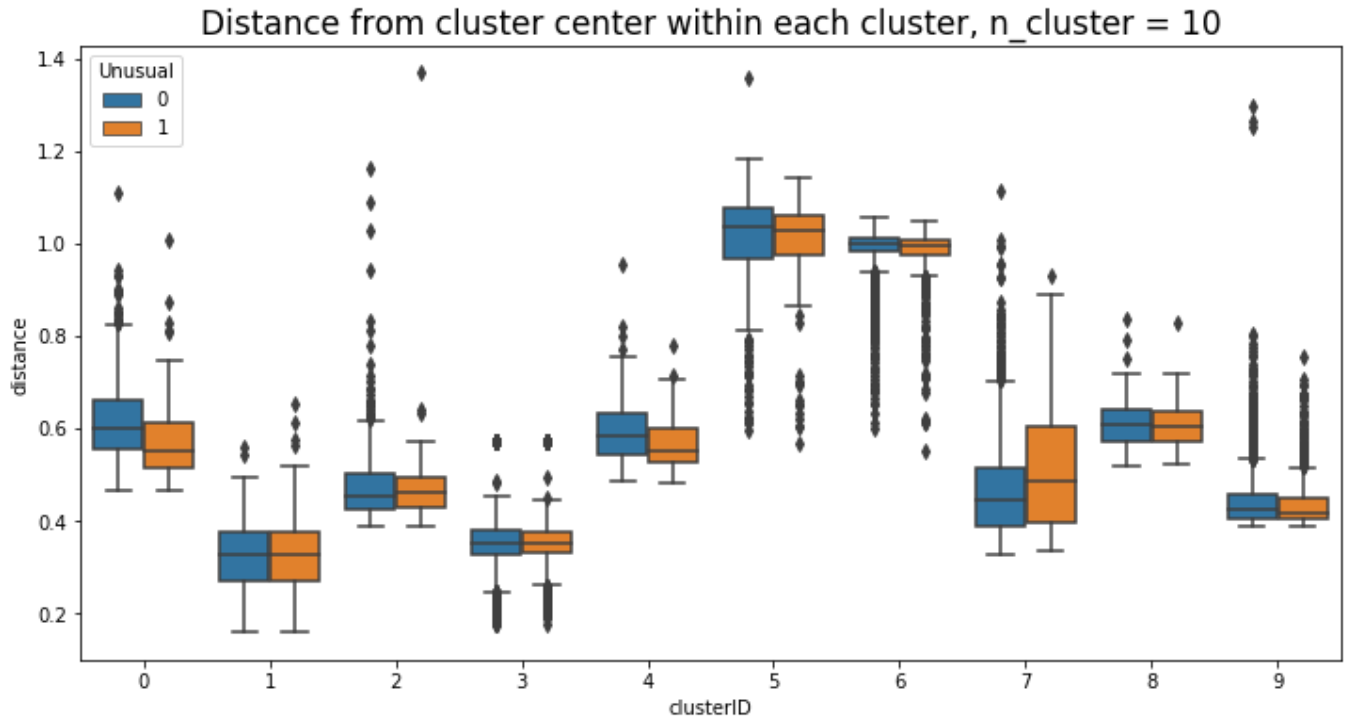


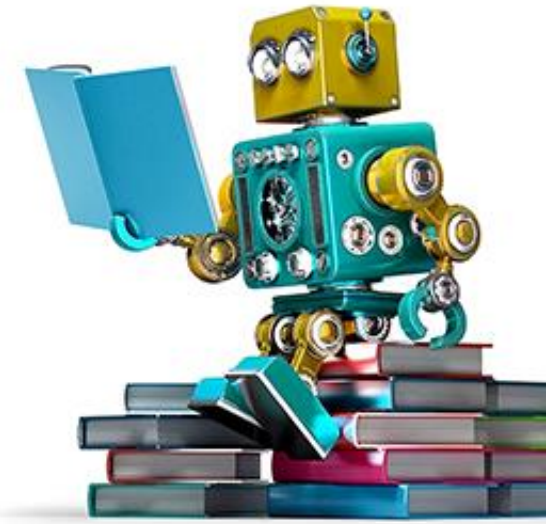
Increasing the n_cluster to 100 resulted in having one cluster with more than 50% prevalence of unusual points



Distance From Cluster Center

- Distance from cluster center for the data points within each cluster is calculated and categorized based on normal and unusual activities
- A function was defined to output average and standard deviation of distance (from cluster center) for unusual point for each cluster



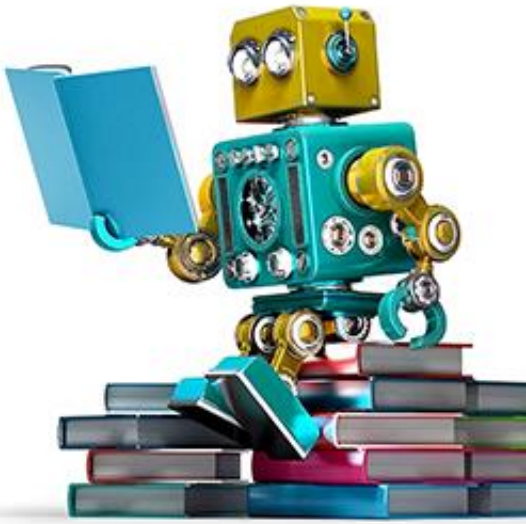


Likelihood of Unforeseen Datapoints Categorized as Unusual Activity

- Any new point will be associated to a cluster based on shortest distance with cluster center points
- The probability of a new datapoint being unusual is defined as:

P = Unusual_prevalence * normal_dist(distance, mu, std)

- Unusual_prevalence is the ratio of the unusual data points to the total data points within each cluster
- normal_dist if probability in normal distribution given mean (mu) and standard deviation (std) of distance (from cluster center) for unusual point for that cluster



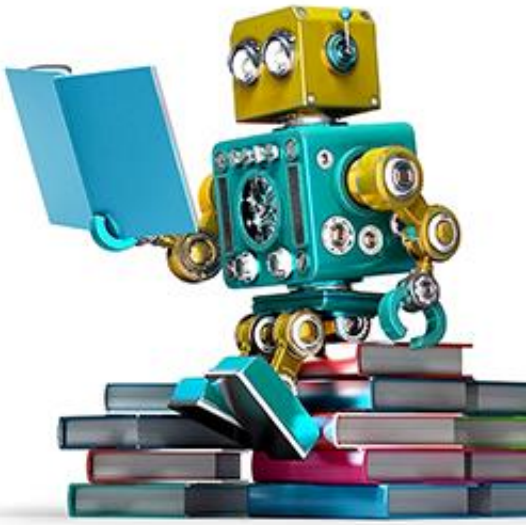
Testing of the Probability Calculating

- The dataset was split to 80% train and 20 test
- A Kmean algorithm with n_cluster = 10 was trained and tested
- Following are the confusion matrix of the results with 1% significance level to identify unusual activities:

	Predcited: Normal	Predcited: Unusual
Actual: Normal	TN = 3699	FP = 1631
Actual: Unusual	FN = 1351	TP = 682

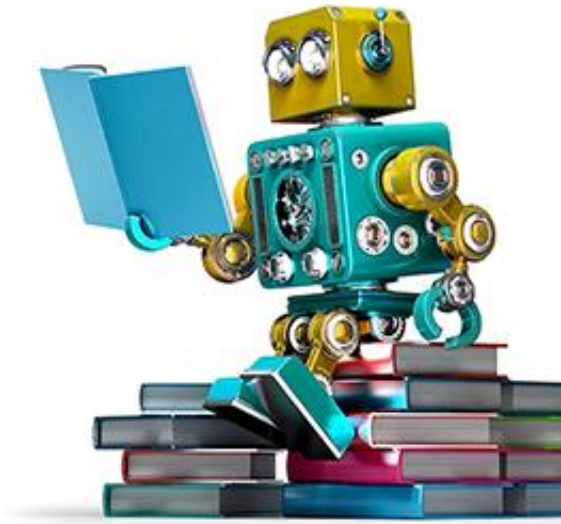
	precision	recall	f1-score	support
Normal	0.732475	0.693996	0.712717	5330.0
Unusual	0.294855	0.335465	0.313852	2033.0
avg/Total	0.611644	0.595002	0.602586	7363.0

A good starting point which needs to be improved



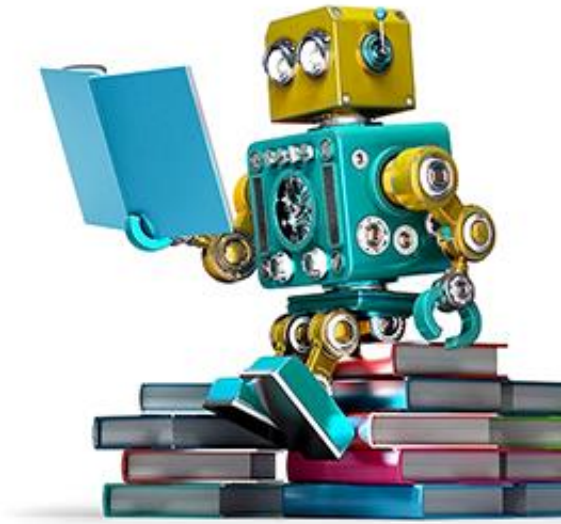
Summary and Conclusion

- Exploratory Data Analysis was conducted to investigate the relationship between different features and visualize distribution of unusual cases
- Multicollinearity was investigated and features with large Variation Inflation Factor (VIF) were dropped from dataset
- Feature Selection was conducted using Z-test and significance level of 0.05 (two tailed test)
- PCA and t-SNE Analysis indicated not clear clusters nor a clearly defined pattern



Summary and Conclusion

- Supervised classification modeling was performed using XGBoost algorithm
  - Parameters optimized for highest recall using a cross-validated grid-search over a parameter grid
  - XGBoost demonstrated exceptional prediction power to classify unusual activities
  - Threshold was adjusted to minimize the FN
  - Optimal threshold was identified using both
    - precision-recall curve (highest F1-score)
    - ROC curve (plot of hit rate vs false alarm rate)



Summary and Conclusion

- Unsupervised anomaly detection conducted using K-means Clustering
  - The prevalence of unusual points in each cluster was investigated to estimate possibility of an unseen point being unusual if being assigned to this cluster
  - Distance of unusual activities points from cluster center was calculated
  - The probability of a new datapoint being unusual is defined as a product:
    - prevalence of unusual points by probability in normal distribution given mean (mu) and standard deviation (std) of distance in each cluster
  - Results with 1% significance level to identify unusual activities on a test set indicated promising results that needs to be improved
  - Comparable results were achieved using both original scaled features and reduced features (4 components of PCA)